

---

# Philosophie der AI

Gerd Graßhoff

2024-06-01

# Table of contents

<b><i>Philosophie der AI</i></b>	<b>15</b>
<b>1 AI-II-1a</b>	<b>16</b>
1.1 Einführung in die Computational Epistemology . . . . .	16
1.1.1 Ein neues Forschungsfeld . . . . .	16
1.1.2 Definition und Abgrenzung . . . . .	16
1.1.3 Zielsetzung und Relevanz . . . . .	17
1.2 Herausforderungen der AI im Bereich des logischen Schließens . . . . .	17
1.2.1 Grundlegende logische Operationen . . . . .	17
1.2.2 Die Debatte um AI und Reasoning . . . . .	17
1.2.3 Historische Perspektiven auf maschinelles Denken . . . . .	17
1.3 Die Zukunft der wissenschaftlichen Forschung durch AI . . . . .	18
1.3.1 Computational AI und Epistemologie . . . . .	19
1.3.2 AI in der Wissenschaft . . . . .	19
1.3.3 AI und Nobelpreise . . . . .	19
1.4 Die Grundlagen moderner AI-Systeme . . . . .	20
1.4.1 Neuronale Netze und Large Language Models . . . . .	20
1.4.2 Sprachkompetenz vs. Sachkompetenz . . . . .	20
1.5 Die Evolution unseres Sprachverständnisses . . . . .	21
1.5.1 Veränderung der Sprachkonventionen . . . . .	21
1.5.2 Erweiterung des Akteursbegriffs . . . . .	21
1.6 Die semantische Revolution in der AI . . . . .	22
1.6.1 Von der Wortübersetzung zur Inhaltsübertragung . . . . .	22
1.6.2 Semantische Inhalte und die epistemologische Wende . . . . .	22
1.7 Die Zukunft der AI: Von der Sprache zur Kognition . . . . .	23
1.7.1 Sprachverarbeitung und akustische Signale . . . . .	23
1.7.2 Von Worten zu Bedeutungen . . . . .	23

1.8	Echtzeit-Übersetzung und ihre Implikationen . . . . .	24
1.8.1	Die nahe Zukunft der Sprachkommunikation . . . . .	24
1.8.2	Ein Gedankenexperiment zur Sprachübersetzung . . . . .	24
1.9	Die Grenzen der Sprache und des Verstehens . . . . .	24
1.9.1	Ausgestorbene Sprachen und künstliche Kommunikation . . . . .	24
1.9.2	Die Grenze zwischen Mensch und Tier in der Kommunikation . . . . .	25
1.10	Die philosophischen Implikationen . . . . .	26
1.10.1	Die Natur der Kommunikation . . . . .	26
1.10.2	Die Verschiebung unserer Wahrnehmung . . . . .	26
1.10.3	Philosophische Fragen zur Natur des Verstehens . . . . .	26
1.11	Empirische Perspektiven auf tierische Kommunikation . . . . .	27
1.11.1	Reiz-Reaktions-Mechanismen vs. Bedeutung . . . . .	27
1.11.2	Die Rolle der Technologie in der Interpretation tierischer Kommunikation . . . . .	27
1.11.3	Der sprechende Papagei - Ein Fallbeispiel . . . . .	28
1.11.4	Herausforderungen der Interpretation . . . . .	28
1.11.5	Die Frage der Bedeutung . . . . .	28
1.11.6	Vergleich zur KI-Debatte . . . . .	29
1.11.7	Ethische und philosophische Implikationen . . . . .	29
<b>2</b>	<b>Was ist AI?</b>	<b>30</b>
2.1	Begrüßung und Einführung . . . . .	30
2.2	AI als Alleskönner . . . . .	32
2.2.1	Der Durchbruch der AI-Visionen . . . . .	32
2.2.2	Die Attraktivität von AI . . . . .	33
2.2.3	Die ursprüngliche Idee des Internets . . . . .	33
2.2.4	Die Ablösung der Webwelt durch AI . . . . .	33
2.2.5	Die Umgestaltung der Architektur des Internets . . . . .	34
2.3	Neue Möglichkeiten durch Künstliche Intelligenz . . . . .	35
2.3.1	Hochwertige Übersetzungen . . . . .	35
2.3.2	Simultanübersetzung und Lektoratsassistentz . . . . .	35
2.3.3	Automatisierte Forschungsberichte . . . . .	36
2.3.4	Das Labor <i>Lettre AI</i> . . . . .	36
2.3.5	Übertragen eines Bildes in maschinenlesbaren Text . . . . .	38
2.3.6	Übersetzen des Textes in eine andere Sprache . . . . .	38

2.4	Erweiterungen . . . . .	38
2.4.1	Analogie zu Sherlock Holmes . . . . .	39
2.4.2	Vielfältige Analysemöglichkeiten von Texten . . . . .	39
2.5	Philosophie als Grundlage für die Möglichkeiten der AI . . . . .	39
2.5.1	Beantwortung von Fragen über Mikrofoneingabe . . . . .	39
2.5.2	Die Möglichkeiten der AI . . . . .	40
2.5.3	Gefahren der AI . . . . .	40
2.5.4	Der sprachliche Kern der AI . . . . .	41
2.5.5	Das Problem der Halluzinationen . . . . .	41
2.5.6	Die Gefahr der Manipulation durch glaubwürdige Fakes . . . . .	42
2.5.7	Selektive Informationen und die Pluralität der Hintergründe . . . . .	42
2.5.8	Die Unausweichlichkeit der AI-Entwicklung und die Notwendigkeit der Gestaltung . . . . .	42
2.5.9	Weitere Gefahren: Diskriminierung und Überwachung . . . . .	43
2.5.10	Die Notwendigkeit der Auseinandersetzung mit AI . . . . .	43
2.6	Nutzungsmöglichkeiten in der Wissenschaft . . . . .	43
2.7	Bislang nicht lösbare Aufgaben . . . . .	44
2.7.1	Frage 1: Einfache Aussage in einer Quelle . . . . .	44
2.7.2	Frage 2: Aussage in Briefen zu einem Thema . . . . .	45
2.7.3	Frage 3: Aussagen einer Person in ihren Schriften . . . . .	45
2.7.4	Frage 4: Keine Aussage einer Person in ihren Schriften . . . . .	45
2.8	Die Herausforderung der inhaltlichen Analyse mit AI . . . . .	46
2.8.1	Grenzen der traditionellen Datenbanken . . . . .	47
2.8.2	Qualifizierte Aussagen auf Basis der verfügbaren Evidenz . . . . .	47
2.8.3	Herausforderungen bei der Interpretation von Metaphern und Ironie . . . . .	47
2.8.4	Lernfähigkeit und Entwicklungspotenzial von AI-Systemen . . . . .	48
2.8.5	Der Paradigmenwechsel durch Large Language Models und Embeddings . . . . .	48
2.8.6	Die Bedeutung der Philosophie für die AI-Forschung . . . . .	49
<b>3</b>	<b>Die Revolution der AI</b>	<b>50</b>
3.1	Begrüßung und Rückblick auf die letzte Vorlesung . . . . .	50
3.2	Traditionell schwer lösbare Fragen in der Forschung . . . . .	50
3.2.1	Noch schwieriger: Evidenz zur Widerlegung von Hypothesen finden . . . . .	50
3.2.2	Komplexe Fragen zur zeitgenössischen Rezeption historischer Hypothesen . . . . .	51
3.3	Die Bedeutung der AI für die Geisteswissenschaften . . . . .	51

---

3.4	Die Evolution der Mensch-Maschine-Interaktion . . . . .	51
3.4.1	Von der Adresseingabe zur Suchanfrage . . . . .	51
3.4.2	Der Durchbruch von Chat-GPT . . . . .	52
3.4.3	Neue Schnittstellen: Sprache, Gesten und Gedanken . . . . .	52
3.5	Die Macht der generativen AI . . . . .	52
3.5.1	Von der Syntax zur Semantik . . . . .	52
3.5.2	Die Bedeutung sprachlicher Ausdrücke . . . . .	52
3.5.3	Philosophische Kritik an der Terminologie . . . . .	53
3.5.4	Von der Zeichenkettensuche zur Bedeutungsanalyse . . . . .	53
3.5.5	Wahrheitswerte und die Welt der Aussagen . . . . .	53
3.5.6	Die Dimension der Aussagen eröffnet neue Möglichkeiten . . . . .	54
3.6	Die drei Säulen der semantischen Revolution . . . . .	54
3.6.1	1. Das Training mit bedeutungsähnlichen Begriffen . . . . .	55
3.6.2	2. Die Frage nach der Bedeutungsähnlichkeit . . . . .	55
3.6.3	3. Das Training mit logischen Regeln . . . . .	56
3.7	Ausblick . . . . .	57
3.7.1	Die Bedeutung von Embeddings . . . . .	57
3.7.2	Die Suche nach bedeutungsähnlichen Aussagen . . . . .	57
3.7.3	Die Erweiterung auf verschiedene Medien . . . . .	58
3.8	Die zweite Revolution: Attention is all you need . . . . .	58
3.8.1	Die Macht der Vorhersage . . . . .	58
3.8.2	Von der Frage zur Anweisung . . . . .	58
3.9	Die Komposition von Instruktionen und Inhalten . . . . .	59
3.9.1	Sprachkompetenz vs. Sachkompetenz . . . . .	59
3.9.2	Gefahren und Grenzen von Chat-GPT . . . . .	60
3.10	Erweiterung der AI-Modelle . . . . .	60
3.10.1	Sachliche Korrektheit und Wahrheit . . . . .	60
3.10.2	Korrespondenztheorie der Wahrheit . . . . .	61
3.11	Sprachentwicklung und Bedeutungsverschiebungen . . . . .	61
3.11.1	Fehltraining und Sprachmarotten . . . . .	61
3.11.2	Reichhaltige Kontextkonstruktion . . . . .	62
3.12	Anwendungsbeispiele und Potenziale . . . . .	62
3.12.1	Übersetzungen als Motor des Trainings . . . . .	62
3.12.2	Zusammenfassungen und Frage-Antwort-Systeme . . . . .	63

<b>4</b>	<b>Charakter von LLMs</b>	<b>64</b>
4.1	Vorlesung Philosophie der AI: Generative Modelle, Large Language Models und Character-Konfiguration . . . . .	64
4.2	Die Revolution der generativen AI-Modelle . . . . .	64
4.2.1	Large Language Models als Kern der generativen AI . . . . .	65
4.2.2	Die Explosion der verfügbaren Modelle . . . . .	65
4.3	Die Funktionsweise der generativen AI-Modelle . . . . .	65
4.3.1	Semantische Ähnlichkeit und Transformation . . . . .	65
4.3.2	Character - Die Formung des künstlichen Charakters . . . . .	65
4.3.3	Metaregeln und kausales Schließen . . . . .	66
4.3.4	Historisches Schließen . . . . .	67
4.3.5	Die Bedeutung des Kontexts . . . . .	67
4.3.6	AGI - Ein umstrittenes Konzept . . . . .	68
4.3.7	Hermeneutik als Herausforderung für AI . . . . .	69
4.3.8	Kontextvergrößerung als Schlüssel zum Verständnis . . . . .	69
4.3.9	Ausblick . . . . .	70
4.3.10	Von der Query zur Instruktion . . . . .	71
4.4	Die Schlüsselemente der Revolution: Semantische Ähnlichkeit und regelhafte Textgenerierung . . . . .	71
4.4.1	Wer war Johann Wolfgang Goethe? - Eine typische Google-Frage . . . . .	71
4.4.2	Die Grenzen der Aktualität . . . . .	72
4.4.3	Interne Präferenzordnungen und Regeln . . . . .	72
4.5	Die Qualität der Internetressourcen reicht nicht aus . . . . .	72
4.5.1	Die Notwendigkeit seriöser Quellen . . . . .	72
4.6	Die Herausforderung: Wahrheit und Wissen . . . . .	73
4.6.1	Der wissenschaftliche Prozess . . . . .	73
4.6.2	Die offene Frage: Der Umgang mit alternativen Lösungen . . . . .	73
4.7	Beispiele zur Veranschaulichung . . . . .	74
4.7.1	Die Macht des Chats . . . . .	74
4.7.2	Kollaborative Intelligenz . . . . .	75
4.8	Herausforderungen und Grenzen aktueller KI-Modelle . . . . .	75
4.8.1	Einstellbare Konversationsstile . . . . .	75
4.8.2	Fragen jenseits von Wikipedia . . . . .	75
4.8.3	Zukünftige Herausforderungen . . . . .	76
4.9	Aktuelle Grenzen und zukünftige Möglichkeiten . . . . .	76

<b>5</b>	<b>LLM für Sprache</b>	<b>78</b>
5.1	Begrüßung und aktueller Stand der AI-Technologie . . . . .	78
5.2	Generative AI und AI-Characters . . . . .	78
5.2.1	Übersetzungsleistung als Beispiel für semantisches Verständnis . . . . .	79
5.3	Herausforderungen und Erwartungen an zukünftige AI-Modelle . . . . .	79
5.3.1	Halluzination als Defizit aktueller Modelle . . . . .	79
5.4	Kompetenzbereiche aktueller und zukünftiger AI-Modelle . . . . .	80
5.4.1	Sprachkompetenz als Basis . . . . .	80
5.4.2	Erweiterbarkeit durch Kontextinformationen . . . . .	80
5.4.3	Bedeutung von Handlungsanweisungen . . . . .	80
5.5	Instruktionsausführung in der AI . . . . .	81
5.6	Lernen von Kompetenz in der AI . . . . .	81
5.6.1	Beispiel: Leonhard Euler . . . . .	82
5.6.2	Weitere Lernmöglichkeiten . . . . .	82
5.6.3	Digitalisierung historischer Bestände . . . . .	83
5.7	Generierung und Kontext in der Interaktion mit Chatmodellen . . . . .	83
5.8	Grenzen aktueller AI-Modelle . . . . .	84
5.9	Erwartungen an eine philosophische AI . . . . .	84
5.9.1	Allgemeine künstliche Intelligenz . . . . .	85
5.10	Semantische Suchen . . . . .	85
5.11	Reasoning . . . . .	86
5.11.1	Charakteristika eines AI-Modells mit Individualität . . . . .	87
5.12	Historische Vorbilder und Metaphern . . . . .	87
5.12.1	Der vitruvische Mensch - Proportion und Harmonie . . . . .	87
5.12.2	David - Freiheit und Selbstbestimmung . . . . .	89
5.12.3	Beuys und der tote Hase - Erklärung und Rechtfertigung . . . . .	91
5.13	Magister AI Faustus - Ein Arbeitstitel für die Zukunft . . . . .	92
5.13.1	Die Zusammenarbeit mit der Klassikstiftung Weimar . . . . .	94
5.13.2	Das Projekt: Goethes Biografie als Herausforderung für AI-Systeme . . . . .	94
5.13.3	Die Vision: Ein erweitertes AI-Modell . . . . .	95
5.14	Die Komplexität der Goethe-Quellen . . . . .	95
5.14.1	Die epistemische Herausforderung . . . . .	96
5.15	Organisation des Projekts . . . . .	96
5.16	Ausblick auf die kommenden Vorlesungen . . . . .	97

<b>6</b>	<b>Sprache und Text</b>	<b>98</b>
6.1	Rückblick auf die letzte Vorlesung . . . . .	98
6.1.1	Vorlesungsmanuskript durch AI generiert . . . . .	98
6.2	Das Projekt “Magister AI Faustus” . . . . .	99
6.2.1	Organisation des Projekts . . . . .	99
6.3	Entwurf einer philosophisch fundierten AI . . . . .	99
6.3.1	Kooperation mit der Klassik Stiftung Weimar . . . . .	100
6.4	Web-Interface unseres AI-Modells . . . . .	100
6.4.1	Logische Beziehungen zwischen Sätzen . . . . .	100
6.4.2	Defizite in logischen Schlussfolgerungen . . . . .	101
6.4.3	Sprachliche Anpassungen ohne Verbesserung . . . . .	102
6.4.4	Lichtblicke und ethische Bedenken . . . . .	102
6.4.5	Notwendigkeit eigener Definitionen . . . . .	102
6.5	Kompetenzen und Grenzen aktueller Modelle . . . . .	102
6.6	Zusammenfassungen generieren . . . . .	103
6.6.1	Technische Details . . . . .	103
6.7	Frage-Antwort-Dialoge . . . . .	104
6.8	Charakteristika und Fähigkeiten der Modelle . . . . .	104
6.8.1	Antwortformate . . . . .	104
6.8.2	Schreibstil . . . . .	104
6.8.3	Fachterminologie . . . . .	105
6.8.4	Kontextbezüge . . . . .	105
6.9	Aktuell bestehende Defizite . . . . .	105
6.9.1	Faktenwissen und logisches Denken . . . . .	105
6.9.2	Hermeneutik und Interpretation . . . . .	105
6.9.3	Kritik und ethische Bewertung . . . . .	106
6.10	Ausblick . . . . .	106
6.10.1	Bewertung der Teilaussagen und Korrekturbedarf . . . . .	107
6.11	Ein philosophisch fundiertes Handlungsmodell für den Umgang mit Instruktionen . . . . .	107
6.11.1	Die zwei Komponenten einer Instruktion als Handlungsanweisung . . . . .	108
6.11.2	Schritte des Handlungsmodells . . . . .	108
<b>7</b>	<b>Denken mit Logik</b>	<b>109</b>
7.1	Begrüßung und Einführung in die 6. Vorlesung . . . . .	109
7.1.1	Stärken und Schwächen der AI-Modelle . . . . .	109

---

7.1.2	Grenzen der AI-Modelle . . . . .	110
7.1.3	Einführung in das Projekt MAGISTER AI Faustus . . . . .	110
7.2	Logisches Denken . . . . .	110
7.2.1	Definition von logischem Denken . . . . .	111
7.2.2	Probleme beim Training von logischem Denken in AI-Modellen . . . . .	111
7.2.3	Unterschied zwischen Context of Discovery und Context of Justification . . . . .	112
7.2.4	Mangel an Discovery-Prozessen in Publikationen . . . . .	112
7.2.5	Experimentelle Untersuchung von Ideen während der Forschung . . . . .	114
7.3	Folgen für AI-Modelle . . . . .	115
7.3.1	Begrenzte Kompetenzen der AI-Modelle . . . . .	116
7.4	Instruktionen für AI-Modelle . . . . .	116
7.4.1	Formulierung von Forschungsvorhaben . . . . .	116
7.4.2	Das Lettre AI Studio . . . . .	117
7.5	Defizite der AI-Modelle bei logischen Verhältnissen . . . . .	117
7.5.1	Zensur bei Anthropic . . . . .	117
7.5.2	OpenAI hat dazugelernt . . . . .	118
7.5.3	Analyse der Sätze durch das einfache Modell . . . . .	118
7.5.4	Konfusion philosophischer Grundfähigkeiten . . . . .	119
7.5.5	Die Anfrage und das Scheitern des Modells . . . . .	120
7.5.6	Verbesserungen und Anpassungen der Modelle . . . . .	120
7.5.7	Die Bedeutung eigener Tests und Erfahrungen . . . . .	120
7.6	Analyse eines verbesserten Modells . . . . .	121
7.6.1	Korrekte Aussagen und logische Verhältnisse . . . . .	121
7.6.2	Problematische Feststellungen und Sachfragen . . . . .	121
7.6.3	Missverständnisse und fehlende Bezüge zur Frage . . . . .	121
7.7	Verbesserung durch Interaktion und Korrektur . . . . .	121
7.7.1	Lernfähigkeit und Grundlagenrevision . . . . .	122
7.8	Ausblick: Philosophie lehrt KI richtiges Denken . . . . .	122
7.9	Die Bedeutung der Schlüssigkeit . . . . .	123
7.9.1	Die Herausforderung der inhaltlichen Suche . . . . .	123
7.10	Die Grenzen aktueller KI-Modelle . . . . .	123
7.10.1	Linguistische Resolution als Lösungsansatz . . . . .	123
7.11	Ein praktisches Beispiel . . . . .	123
7.11.1	Analyse des Arguments . . . . .	124
7.11.2	Das Wahrheitswerttafelverfahren . . . . .	124

7.12	Die Macht der erweiterten Instruktionen . . . . .	124
7.13	Fazit . . . . .	125
<b>8</b>	<b>Argumente</b>	<b>126</b>
8.1	Begrüßung und Einleitung . . . . .	126
8.1.1	Rasante Entwicklung der KI-Technologie . . . . .	126
8.2	Sprachverarbeitung durch KI . . . . .	127
8.2.1	Allgemeingültigkeit des Beispiels . . . . .	127
8.2.2	Große Dokumentenbestände . . . . .	128
8.2.3	Begrenzte Nutzbarkeit derzeitiger Digitalisate . . . . .	128
8.2.4	Philosophische Konzepte hinter der KI-Sprachverarbeitung . . . . .	128
8.2.5	Verarbeitung durch KI-Modelle . . . . .	129
8.2.6	Verhalten von Chat-GPT . . . . .	129
8.2.7	Nutzerseitige Steuerungsmöglichkeiten . . . . .	130
8.3	Erste Befunde KI-Modelle . . . . .	130
8.3.1	Grenzen aktueller AI-Modelle bei der Informationsverarbeitung . . . . .	131
8.3.2	Philosophische Ansätze zur Verbesserung von AI-Modellen . . . . .	131
8.3.3	Identifikation und Beurteilung von Argumenten als zentrale Aufgabe der AI . . . . .	132
8.3.4	Philosophische Befunde zur Argumentation . . . . .	132
8.3.5	Unterstützung von AI-Modellen bei der Argumentanalyse . . . . .	132
8.3.6	Ein Schema der Argumentanalyse . . . . .	133
8.3.7	Die Kernfrage . . . . .	134
8.3.8	Zuverlässigkeit der Argumentanalyse durch AI . . . . .	134
8.4	Ein Beispiel für Argumentanalyse . . . . .	134
8.4.1	Schlüssigkeit vs. Triftigkeit . . . . .	135
8.4.2	Methodisches Vorgehen bei der Argumentanalyse . . . . .	135
8.4.3	Ein Verfahren nach Wittgensteins Tractatus . . . . .	136
8.4.4	Anwendung der Methode auf ein konkretes Beispiel . . . . .	136
8.5	Analyse mit Wahrheitstabellen . . . . .	136
8.5.1	Identifikation der Aussagen im Text . . . . .	137
8.5.2	Berechnung der Wahrheitswerte . . . . .	137
8.5.3	Fehleranalyse und Korrektur . . . . .	137
8.5.4	Ausblick auf weiterführende Methoden . . . . .	138

<b>9 Akteure</b>	<b>140</b>
9.1 Was ist ein AI-Akteur und wie wird er gebaut? . . . . .	140
9.1.1 Sprach- und Bedeutungskompetenz als Basis . . . . .	140
9.1.2 Wissen, Kompetenzen und Charakter . . . . .	141
9.2 Apples Umsetzung von AI-Akteuren . . . . .	141
9.2.1 Multimediale Verarbeitung und tiefe Integration . . . . .	141
9.2.2 Datenschutz und Privatsphäre . . . . .	142
9.3 Philosophische Konzepte hinter AI-Akteuren . . . . .	142
9.3.1 Informationsquellen und Handlungsmodell . . . . .	143
9.3.2 Zerlegung des Textes in Einzelsätze und Standardisierung . . . . .	145
9.3.3 Zuweisung von Buchstaben zu Aussagen gleicher Bedeutung . . . . .	145
9.3.4 Umsetzung des Arguments in eine kompakte Form . . . . .	146
9.3.5 Anwendung von Wahrheitstabellen zur Überprüfung der Schlüssigkeit . . . . .	146
9.4 Demonstration des Verfahrens mit ChatGPT 4.0 . . . . .	146
9.4.1 Beispieltext und Fragestellung . . . . .	147
9.4.2 Anforderungen an AI-Systeme . . . . .	147
9.5 Interaktion mit dem AI-Agenten . . . . .	147
9.5.1 Nutzung von ChatGPT und Antwortzeiten . . . . .	148
9.5.2 Der AI-Agent als Diskussionspartner . . . . .	148
9.5.3 Verarbeitung von Bildern und Argumenten . . . . .	148
9.5.4 Bilderkennung und Wissensgenerierung . . . . .	148
9.5.5 Extraktion und Verarbeitung von Textinformationen aus Bildern . . . . .	149
9.5.6 Erläuterung der einzelnen Schritte . . . . .	149
9.5.7 Erstellung der Wahrheitstabelle . . . . .	150
9.6 Das Akteurmodell als Grundlage der Interaktion mit AI . . . . .	150
9.6.1 Die Bedeutung von Instruktionstexten in der Wissenschaft . . . . .	151
9.7 Die Anwendung des Instruktionsschemas auf AI-Akteure . . . . .	151
9.7.1 Variationsmöglichkeiten der Instruktionen . . . . .	151
9.7.2 Entwicklungspotenzial der Nutzungsschnittstellen . . . . .	152
9.8 Ergebnisse und Perspektiven . . . . .	152
9.8.1 Perspektiven für den Einsatz von AI . . . . .	152
9.8.2 Ausblick auf die Entwicklung von Instruktionsbibliotheken . . . . .	153
9.8.3 Perspektiven für kollaborative Forschung . . . . .	153
9.9 Auswertung und Validitätsprüfung . . . . .	153
9.9.1 Ablauf der Planausführung . . . . .	154

9.10	Standardisierung der Instruktion . . . . .	154
9.10.1	Formanalyse und Inferenz . . . . .	154
9.10.2	Epistemische Inferenz . . . . .	155
9.11	Logische Verfahren . . . . .	155
9.11.1	Inferenzanalyse . . . . .	155
9.12	Epistemische Inferenz und Hermeneutik . . . . .	156
9.12.1	Erschließung alter Sprachen . . . . .	156
9.13	Analyse der Wissenschaftsliteratur . . . . .	157
9.14	Das Handlungsmodell . . . . .	157
<b>10</b>	<b>Validierung</b>	<b>158</b>
10.1	Einleitung und Rückblick . . . . .	158
10.1.1	Kompetenzen der AI-Akteure . . . . .	158
10.1.2	Erweiterungen der Kernkompetenzen . . . . .	158
10.2	User Interface - Die Schnittstelle zwischen Mensch und Maschine . . . . .	159
10.2.1	Expansion der Sprachsteuerung . . . . .	159
10.2.2	Aufstieg der AI-Bots . . . . .	159
10.3	Bedeutung von Wissensquellen . . . . .	160
10.3.1	GitHub: Eine Schatzkammer für KI-Modelle . . . . .	160
10.3.2	Von Fehlerdiskussionen zu Syntax-Änderungen . . . . .	161
10.3.3	Die Macht der Programmiersprachenkenntnisse . . . . .	161
10.4	Neue Informationsquellen für KI-Modelle . . . . .	161
10.4.1	Websuche: Aktuelles Wissen ergänzt historische Daten . . . . .	161
10.4.2	Informationsbroker: Spezialisierte und verlässliche Datenquellen . . . . .	161
10.5	Herausforderungen und Grenzen von KI-Modellen . . . . .	162
10.5.1	Rechtssammlung: Ein Beispiel für Interpretationsschwierigkeiten . . . . .	162
10.5.2	Halluzinationen und fehlende kritische Hinterfragung . . . . .	162
10.5.3	Mögliche negative Folgen und Regulierungsbedarf . . . . .	163
10.6	Ausblick: Weitere Anwendungsbereiche und RAG . . . . .	163
10.6.1	Informationsbroker in verschiedenen Bereichen . . . . .	163
10.6.2	Retrieval Augmented Generation (RAG) . . . . .	163
10.6.3	Die RAG-Komponente . . . . .	164
10.6.4	Die Dimensions-Datenbank . . . . .	164
10.6.5	Der digitalisierte Wissensschatz . . . . .	165
10.6.6	Zugang zu wissenschaftlichen Publikationen . . . . .	166

10.6.7	Die Wahrheitsfrage . . . . .	166
10.6.8	Überprüfung einer Aussage zu GPT-3 Parametern . . . . .	167
10.6.9	Grenzen der AI bei philosophischen und geisteswissenschaftlichen Anfragen . . . . .	168
10.7	Bewertung der Leistungsfähigkeit und Grenzen der AI . . . . .	168
10.7.1	Imitation einer Rechtfertigung statt echter Begründung . . . . .	168
10.7.2	Fehlende Berücksichtigung von Gegenargumenten und alternativen Sichtweisen . . . . .	168
10.7.3	Potenzial für die Untersuchung kontroverser Studien . . . . .	168
10.8	Ausblick auf zukünftige Entwicklungen . . . . .	169
10.8.1	Anforderungen an eine epistemische Bewertung durch AI . . . . .	169
10.8.2	Bausteine für die zukünftige Forschung . . . . .	169
<b>11</b>	<b>Epistemische Bewertung</b>	<b>170</b>
11.1	Die philosophische Vogelperspektive auf AI-Modelle . . . . .	170
11.1.1	Dimensionen der epistemischen Bewertung . . . . .	170
11.1.2	Ein herausforderndes Puzzle für AI-Modelle . . . . .	170
11.2	Analyse der AI-Antworten . . . . .	171
11.2.1	Anthropic Cloud Haiku's Antwort . . . . .	171
11.2.2	Das Kernproblem der epistemischen Dimension . . . . .	171
11.3	Die Grenzen aktueller AI-Modelle . . . . .	171
11.3.1	Methodische Fehler im Reasoning-Apparat . . . . .	172
11.4	Verbesserungen und neue Ansätze . . . . .	172
11.4.1	Claude Sonnet's Lösungsansatz . . . . .	172
11.4.2	Vergleich mit anderen Modellen . . . . .	172
11.5	Die Herausforderung der epistemischen Sicherheit . . . . .	173
11.5.1	Ein komplexeres Beispiel: Goethe und das Kartenspiel . . . . .	173
11.5.2	Die KI-generierte Antwort und ihre Schwächen . . . . .	173
11.5.3	Die Notwendigkeit einer fundierten Quellenprüfung . . . . .	174
11.5.4	Ein konkretes Beispiel: Goethes Brief von 1773 . . . . .	174
11.5.5	Die Herausforderungen bei der AI-gestützten Analyse historischer Quellen . . . . .	174
11.5.6	Die Grenzen aktueller AI-Systeme . . . . .	174
11.6	Ausblick: Die Zukunft der AI in der Wahrheitsfindung . . . . .	175
11.6.1	Das Kernproblem der Falschinformationen . . . . .	176
11.6.2	Die Beweislast-Problematik . . . . .	176
11.7	Strategien zur Verbesserung der AI-Antworten . . . . .	176
11.7.1	Der Ansatz von Anthropic . . . . .	176

---

11.7.2 Die Bedeutung von Bildinformationen . . . . .	177
11.8 Die Anforderungen an eine zuverlässige AI . . . . .	177
11.8.1 Evidenzbasierte Antworten . . . . .	177
11.8.2 Nachweisbarkeit und Rechtfertigung . . . . .	177
11.9 Die Herausforderungen der historischen Beweisführung . . . . .	178
11.9.1 Komplexität einfacher Fragen . . . . .	178
11.9.2 Die Rolle der Computertechnologie . . . . .	178
11.9.3 Zerlegung in Teilaufgaben . . . . .	178
11.10 Die philosophische Dimension . . . . .	179
11.10.1 Universelle Anwendbarkeit . . . . .	179
11.10.2 Die Lösung des Halluzinationsproblems . . . . .	179
11.11 Die Bedeutung von Kulturinstitutionen . . . . .	179
11.11.1 Der Wert des Quellenmaterials . . . . .	179
11.11.2 Konkrete Beispiele . . . . .	179
11.11.3 Multimodale Fähigkeiten der AI . . . . .	180
11.12 Die Natur von Aussagen und ihre Bedeutung . . . . .	180
11.12.1 Aussagen als abstrakte Entitäten . . . . .	180
11.12.2 Die Vielfalt sprachlicher Ausdrucksformen . . . . .	181
11.13 Die Komplexität scheinbar einfacher Fragen . . . . .	181
11.13.1 Die Frage nach Goethes Kartenspiel . . . . .	181
11.13.2 Analyse der Schlüsselbegriffe . . . . .	181
11.13.3 Wahrheitsbedingungen und Evidenz . . . . .	181
11.14 AI-Modelle und ihre Grenzen . . . . .	182
11.14.1 Halluzinationen und ihre Ursachen . . . . .	182
11.14.2 Der Weg zur Verbesserung . . . . .	182
11.15 Die Rolle des Kontexts und der Expertise . . . . .	182
11.15.1 Die Notwendigkeit von Fachkenntnissen . . . . .	182
11.15.2 Komprimierung von Expertise . . . . .	182
11.16 Hermeneutik und AI . . . . .	183
11.16.1 Die Herausforderung der Bedeutungsinterpretation . . . . .	183
11.16.2 Ein Lösungsansatz: Standardreferenzen . . . . .	183
11.16.3 Kriterien für Bedeutungsinterpretation . . . . .	183
11.17 Schlussbetrachtungen zur epistemischen Bewertung . . . . .	183

---

<b>12 Instruktion</b>	<b>184</b>
12.1 Grundelemente der generativen AI . . . . .	184
12.1.1 Die Komplexität des Inputs . . . . .	184
12.1.2 Die Verarbeitungskapazität moderner Modelle . . . . .	184
12.1.3 Die Entwicklung der Modellkapazität . . . . .	185
12.1.4 Begrenzungen der Ausgabe . . . . .	185
12.1.5 Praktische Konsequenzen der Begrenzungen . . . . .	185
12.2 Herausforderungen bei der Formulierung von Instruktionen . . . . .	186
12.2.1 Die Vielfalt der Eingabeformate . . . . .	186
12.2.2 Der Prozess der Instruktionsverarbeitung . . . . .	186
12.3 Ein konkretes Beispiel: Die Fontänen von Sanssouci . . . . .	187
12.3.1 Die irreführende Antwort der AI . . . . .	187
12.3.2 Die historische Wahrheit . . . . .	187
12.3.3 Die Herausforderung für die AI . . . . .	187
12.3.4 Die Konsequenzen fehlerhafter Antworten . . . . .	188
12.4 Die Suche nach Lösungen . . . . .	188
12.4.1 Aktuelle Forschungsansätze . . . . .	188
12.4.2 Die Notwendigkeit epistemischer Expertise . . . . .	188
12.4.3 Die Rolle des Prompts . . . . .	188
12.4.4 Der Weg nach vorn . . . . .	188
12.4.5 Das Problem der Faktentreue . . . . .	189
12.4.6 Die Notwendigkeit kritischer Überprüfung . . . . .	189
12.4.7 Die Rolle des menschlichen Inputs . . . . .	190
12.5 Strategien zur Verbesserung der AI-Antworten . . . . .	190
12.5.1 Präzisere Prompts und Nachfragen . . . . .	190
12.5.2 Entwicklung von AI-gestützten Publikationsassistenten . . . . .	190
12.5.3 Integration von Wahrheitsüberprüfungen . . . . .	191
12.6 Potenziale und Grenzen der AI in der Wissenschaft . . . . .	191
12.6.1 Der Fall der kalten Kernfusion . . . . .	191
12.6.2 Chancen und Risiken . . . . .	192
12.7 Ausblick: Die Zukunft der AI in der Wissenschaft . . . . .	192
<b>References</b>	<b>194</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 AI-II-1a

## 1.1 Einführung in die Computational Epistemology

Willkommen zur zweiten Vorlesung über die Philosophie der Künstlichen Intelligenz. Obwohl diese als Teil 2 angekündigt wurde, möchte ich betonen, dass der erste Teil keine zwingende Voraussetzung für das Verständnis dieser Vorlesung darstellt. Heute werden wir uns einem spezifischen und höchst relevanten Themenfeld widmen, das ich als “Computational Epistemology” bezeichne.

### 1.1.1 Ein neues Forschungsfeld

Mit der Computational Epistemology beabsichtige ich, ein völlig neues Gebiet für die Philosophie zu erschließen und zu etablieren. Nach meinem derzeitigen Kenntnisstand existiert dieses Feld in dieser Form noch nicht, obwohl es naheliegend erscheint, dass ähnliche Ideen weltweit bereits aufgekommen sein könnten. Die Einzigartigkeit dieses Ansatzes liegt in seinen neuartigen Methoden und Zielsetzungen, gepaart mit einer universellen Bedeutung.

### 1.1.2 Definition und Abgrenzung

Epistemologie, abgeleitet vom griechischen Wort “Episteme” für Wissen, befasst sich mit der Lehre vom Wissen oder der Erkenntnistheorie. Die Computational Epistemology zielt darauf ab, diese philosophische Disziplin mit den Möglichkeiten und Herausforderungen der modernen AI zu verbinden. Dabei geht es nicht darum, die gesamte philosophische Erkenntnistheorie zu digitalisieren, sondern vielmehr darum, jene Aspekte zu selektieren und anzuwenden, die für die Entwicklung und das Verständnis von AI-Modellen besonders relevant sind.

### **1.1.3 Zielsetzung und Relevanz**

Mein Anspruch ist es, die Erkenntnistheorie in den Dienst der Artificial Intelligence zu stellen, um deren spezifische Bedürfnisse und Anforderungen zu adressieren. Im Laufe dieser Vorlesung werde ich aufzeigen, dass die derzeit existierenden AI-Modelle erhebliche Lücken und teilweise fehlerhafte Annäherungen an epistemologische Fragen aufweisen. Diese Defizite zu beheben, stellt ein enormes Potenzial für die Weiterentwicklung der AI dar.

## **1.2 Herausforderungen der AI im Bereich des logischen Schließens**

### **1.2.1 Grundlegende logische Operationen**

In der vorherigen Vorlesung haben wir uns mit einigen elementaren Beispielen der Logik befasst, insbesondere mit den Regeln des logischen Schließens. Diese Fähigkeit, aus einer gegebenen Menge von Prämissen deduktiv Schlüsse zu ziehen, ist eine fundamentale Übung in der Philosophie, die bereits in den ersten Wochen des Studiums behandelt wird. Überraschenderweise haben selbst fortschrittliche AI-Modelle noch Schwierigkeiten, diese scheinbar einfachen logischen Operationen zuverlässig durchzuführen.

### **1.2.2 Die Debatte um AI und Reasoning**

In den sozialen Medien und der wissenschaftlichen Gemeinschaft entbrennt regelmäßig eine hitzige Diskussion darüber, ob AI-Systeme überhaupt in der Lage sind, logische Schlussfolgerungen zu ziehen. Einige Kritiker behaupten sogar, dass dies prinzipiell unmöglich sei. Die Argumente reichen von der Notwendigkeit des Verstehens für das Schließen bis hin zu fundamentalen Unterschieden zwischen menschlicher und maschineller Kognition.

### **1.2.3 Historische Perspektiven auf maschinelles Denken**

Um diese Debatte in einen breiteren Kontext zu setzen, lohnt sich ein Blick in die Geschichte. Es gibt faszinierende Beispiele für mechanische Systeme, die komplexe Berechnungen und Vorhersagen treffen konnten, lange bevor der Begriff der Künstlichen Intelligenz geprägt wurde.

### **1.2.3.1 Der Antikythera-Mechanismus**

Ein besonders eindrucksvolles Beispiel ist der Antikythera-Mechanismus, ein hochkomplexes Räderwerk aus Bronze, das im 2. Jahrhundert vor Christus konstruiert wurde. Dieses erstaunliche Gerät war in der Lage, Sonnen- und Mondfinsternisse mit höchster Präzision vorherzusagen - eine Fähigkeit, die für politische und militärische Entscheidungen von enormer Bedeutung war.

### **1.2.3.2 Wissenschaftliche Kompetenz im Altertum**

Die Existenz solcher Geräte wirft interessante Fragen über das Wesen des Schließens und der Erkenntnisgewinnung auf. In der Antike war die Fähigkeit, astronomische Ereignisse vorherzusagen, nicht an das individuelle Verständnis eines Menschen gebunden, sondern an die Expertise spezialisierter Zentren wie der Bibliotheken in Babylon. Diese Zentren bewahrten und pflegten das notwendige Wissen über Jahrhunderte hinweg, unabhängig von politischen Umwälzungen.

### **1.2.3.3 Reasoning als unabhängiges Konzept**

Diese historischen Beispiele zeigen, dass das Konzept des "Reasoning" - des vernunftgeleiteten Schließens - nicht zwangsläufig an menschliche Einsicht gebunden sein muss. Vielmehr geht es um die Existenz und Anwendung logischer Zusammenhänge zwischen verschiedenen Sachverhalten und deren Konsequenzen. Der Antikythera-Mechanismus demonstriert eindrucksvoll, wie komplexe Berechnungen und Vorhersagen durch mechanische Systeme realisiert werden können, ohne dass ein tiefgreifendes theoretisches Verständnis bei jedem Anwender vorausgesetzt werden muss.

## **1.3 Die Zukunft der wissenschaftlichen Forschung durch AI**

Die historischen Beispiele führen uns zu einer hochaktuellen Debatte: Wie lange wird es dauern, bis wir Maschinen entwickeln, die in der Lage sind, wissenschaftliche Forschung auf einem Niveau zu betreiben, das dem der besten menschlichen Spezialisten ebenbürtig oder gar überlegen ist? Diese Frage beschäftigt nicht nur die Fachwelt, sondern findet auch in renommierten Publikationen wie der Financial Times und dem Time Magazine Beachtung.

Die Meinungen und Prognosen gehen weit auseinander. Einige Experten argumentieren, dass es prinzipiell unmöglich sei, während andere konkrete Zeiträume für diesen Durchbruch vorhersagen. Diese Diskus-

sion erinnert stark an die Debatten, die schon vor Jahrtausenden über die Fähigkeiten mechanischer Systeme geführt wurden.## Die revolutionäre Entwicklung der AI

In den letzten Jahren haben wir eine faszinierende Entwicklung im Bereich der Künstlichen Intelligenz, oder wie wir sie fortan nennen werden, AI, miterlebt. Wir befinden uns aktuell in einer äußerst spannenden Phase, in der das kreative Potenzial der AI immer deutlicher zutage tritt. Es ist davon auszugehen, dass wir bereits Ende nächsten Jahres einen bedeutenden Meilenstein erreichen werden.

In dieser Vorlesung möchte ich Sie durch die faszinierenden Mechanismen führen, die wir entschlüsseln und maschinell umsetzen, um die Prozesse des Wissenserwerbs, der Wissenssuche und der Wissenskritik zu verstehen und nachzubilden. Wir werden uns damit beschäftigen, wie AI-Systeme die notwendigen Operationen durchführen, um Antworten zu generieren, die unseren Wissensansprüchen genügen.

### **1.3.1 Computational AI und Epistemologie**

Unser Ziel ist es, im Rahmen der Computational AI und Computational Epistemologie zu ergründen, welche funktionalen Komponenten bereitgestellt werden müssen, um diese komplexen Aufgaben zu bewältigen. Auch wenn wir möglicherweise nicht alle Antworten bis zum Ende dieser Vorlesung finden werden, bin ich überzeugt, dass wir bis Ende nächsten Jahres bereits erhebliche Fortschritte gemacht haben werden.

### **1.3.2 AI in der Wissenschaft**

Lassen Sie mich eine gewagte Prognose wagen: Ich bin fest davon überzeugt, dass wir in wenigen Jahren, sicherlich vor 2030, Maschinen haben werden, die auf den speziellsten und kompliziertesten Feldern der Wissenschaft eigenständig kreative Leistungen erbringen werden. Ob sie dies alleine oder in Kooperation mit Menschen tun werden, ist dabei zweitrangig. Sie werden zweifellos wie hochqualifizierte Kollegen agieren.

### **1.3.3 AI und Nobelpreise**

Ein bemerkenswerter Indikator für diese Entwicklung war die diesjährige Vergabe der Nobelpreise. Betrachten wir die verschiedenen Disziplinen, so fällt auf, dass fast jede den Nobelpreis an Arbeiten verliehen hat, die entweder mithilfe von AI erstellt wurden oder sich mit den Grundlagen von AI befassten. Dies gilt

sogar für die Ökonomie, wo einer der Preisträger vom MIT für seine Forschung über die wirtschaftlichen Auswirkungen von AI ausgezeichnet wurde.

Interessanterweise hatte das AI-Modell Claude nur eine Woche zuvor diesen Preisträger korrekt vorhergesagt - ein bemerkenswerter Treffer, der die Leistungsfähigkeit moderner AI-Systeme unterstreicht. Dies markiert einen bedeutenden Wendepunkt: Seit Herbert Simon in den 1970er Jahren ist dies das erste Mal im 21. Jahrhundert, dass Forscher im Zusammenhang mit AI-Forschung in ihren jeweiligen Gebieten den Nobelpreis erhalten haben.

## **1.4 Die Grundlagen moderner AI-Systeme**

### **1.4.1 Neuronale Netze und Large Language Models**

Die Grundlagen für die heutigen AI-Systeme wurden bereits vor Jahrzehnten gelegt. Der diesjährige Physik-Nobelpreis würdigte die theoretische Formulierung neuronaler Netze, die als technische Imitate der Hirnfunktion fungieren. Obwohl diese frühen Modelle nur entfernt mit den heutigen Implementierungen zu tun haben, legten sie den evolutionären und technologischen Grundstein für die Fähigkeiten moderner Modelle.

Die heutigen Large Language Models (LLMs) und Transformationsmodelle gehen weit über diese ursprünglichen Konzepte hinaus. Eine zentrale Frage ist, ob wir weiterhin massive Ressourcen in die Entwicklung und technische Umsetzung solcher Modelle investieren müssen, um bis 2030 wissenschaftlich kreative Kompetenzen und Intelligenz zu generieren. Meine Antwort darauf ist ein klares Nein, und ich werde im Verlauf dieser Vorlesung erläutern, warum.

### **1.4.2 Sprachkompetenz vs. Sachkompetenz**

Die aktuellen AI-Modelle, die Sie vielleicht von ChatGPT oder ähnlichen Anwendungen kennen, sind im Grunde genommen Sprachmodelle. Sie verfügen über eine beeindruckende Sprachkompetenz, die sie durch die Verarbeitung enormer Mengen sprachlichen Wissens aus verschiedensten Bereichen erworben haben. Dies erweckt oft den Eindruck, dass sie auch über Sachkompetenz verfügen - doch das ist ein Trugschluss.

Lassen Sie mich dies mit einem Vergleich verdeutlichen: Stellen Sie sich einen fortschrittlichen Taschenrechner vor, der ein Display hat, auf dem Berechnungen angezeigt werden. Je besser dieser Taschenrechner programmiert ist, desto überzeugender ist der Zusammenhang zwischen Eingabe, Rechenopera-

tion und Ausgabe. Wenn Sie “3 + 5” eingeben und das Display “8” anzeigt, könnten Sie geneigt sein, die Maschine für intelligent zu halten. Doch ist sie wirklich intelligent, oder ist es vielmehr der Bau dieser Maschine, der intelligent ist?

## **1.5 Die Evolution unseres Sprachverständnisses**

### **1.5.1 Veränderung der Sprachkonventionen**

Interessanterweise führt die Entwicklung von AI-Systemen dazu, dass sich unser Sprachverständnis und unsere Beschreibung dessen, was in diesen Systemen geschieht, allmählich verändert. Lassen Sie mich dies an einem Beispiel verdeutlichen:

Vor 20 Jahren hätte die Aufforderung, die dritte Quadratwurzel von 1000 zu berechnen, wahrscheinlich zu zahlreichen Handmeldungen geführt. Heute würden die meisten von uns reflexartig zum Taschenrechner greifen. Wenn ich Sie nun fragen würde, was Ihr Rechner ausrechnet, würden Sie das vermutlich für einen völlig normalen Satz halten.

Doch beachten Sie, wie sorglos wir mittlerweile mit dem Verb “ausrechnen” umgehen. Wir formulieren es so, als ob der Rechner derjenige wäre, der etwas ausrechnet. Vor zwei Jahrzehnten wäre das als Kategorienfehler betrachtet worden. Die natürliche Redeweise wäre gewesen: “Sie als Person rechnen etwas mit Hilfe des Geräts aus.”

### **1.5.2 Erweiterung des Akteursbegriffs**

Heute hat sich die Zulässigkeit der Beschreibung des Akteurs erweitert. Wir haben kein Problem mehr damit, den Akteursbegriff von dem zwingend handelnden menschlichen Akteur auf Instrumente auszuweiten. Dies zeigt sich auch in anderen Bereichen: Während es früher undenkbar erschien, dass ein Computer “lesen” könnte, akzeptieren wir heute problemlos Formulierungen wie “Der Computer liest eine Diskette oder einen USB-Stick”.

Diese Verschiebung in der Epistemologie der Verben kognitiver Vermögen ist bemerkenswert. Sie verdeutlicht, warum grundsätzliche Debatten darüber, ob Computer oder AI dies oder jenes können, oft fruchtlos sind. Am Ende läuft es darauf hinaus festzustellen, dass für den einen der Begriff des Lesens die Existenz einer vernunftgeleiteten Person voraussetzt, während der andere ihn völlig personenunabhängig verwendet.

## **1.6 Die semantische Revolution in der AI**

### **1.6.1 Von der Wortübersetzung zur Inhaltsübertragung**

Ein entscheidender Durchbruch in der AI-Forschung erfolgte vor etwa zehn Jahren im Kontext der Google Translate Forschungsgruppe. Ihre Aufgabe bestand darin, aus bestehenden Texten in einer Sprache A Texte in einer beliebigen anderen Sprache zu generieren. In einem bahnbrechenden Paper wurde dieser Übersetzungsprozess auf ein fundamentales Prinzip reduziert: Die Modelle werden so lange trainiert, bis sie die Abfolge von Übersetzungswörtern, die aus dem Ursprungssatz entnommen wurden, korrekt und sinngemäß als Ausgabe erzeugen können.

Der entscheidende Fortschritt gegenüber früheren, oft lächerlichen Wort-für-Wort-Übersetzungen besteht darin, dass die neuen Modelle in der Lage sind, die Komplexität des Kontexts zu erfassen. Dies führt dazu, dass die Ausgabesätze nicht nur grammatikalisch korrekt sind, sondern auch die richtigen Wortbedeutungen im jeweiligen Kontext wählen.

### **1.6.2 Semantische Inhalte und die epistemologische Wende**

Das Training dieser Modelle wurde so konzipiert, dass sie inhaltlich invariant sein sollten - die Inhalte sollten durch die Übersetzung nicht verändert werden. Diese Anforderung an die Übersetzung markiert den Beginn der AI-Revolution. Es geht nun darum, Modelle zu trainieren, die semantische Inhalte identifizieren können. Sie können entscheiden, ob zwei Ausdrücke inhaltlich näher zueinander stehen als ein dritter.

Diese semantische Revolution besteht darin, dass wir nun über Inhalte sprechen können und nicht mehr nur über Ausdrücke. Damit betreten wir zum ersten Mal das Gebiet der Epistemologie. Wir können nun Aussagen über Aussagen treffen. Die Epistemologie, die Logik und das Wissen haben weniger mit Sätzen zu tun als mit den Inhalten, die diese Sätze ausdrücken. Erst seit diesem großen Schritt in der Entwicklung von Übersetzungsprogrammen haben wir in der AI Zugang zu dieser Ebene der Inhalte erhalten.

## 1.7 Die Zukunft der AI: Von der Sprache zur Kognition

### 1.7.1 Sprachverarbeitung und akustische Signale

Die aktuellen Sprachmodelle haben eine zusätzliche Kompetenz erworben: Sie können sprachliche Signale verarbeiten. Stellen Sie sich vor, wie mein gesprochener Vortrag in einem akustischen Spektrogramm aussieht - eine wellenartige Zusammenstellung, die Lautstärke, Tonhöhe und andere akustische Eigenschaften modelliert. Die AI-Systeme sind nun in der Lage, diese komplexen Tonsignale zu erfassen und zu interpretieren.

Die Aufgabe besteht darin, die Zusammensetzung dieser akustischen Signale bestimmten Wörtern zuzuordnen - im Grunde eine Klassifikationsaufgabe. Vor zwanzig Jahren funktionierte dies noch recht unzuverlässig, aber in den letzten zehn Jahren haben wir enorme Fortschritte gemacht. Denken Sie nur an die Sprachfunktion Ihres Handys und wie gering die Fehlerquote heutzutage ist - und das ohne einen Großrechner im Hintergrund.

Diese Entwicklungen zeigen, dass wir an der Schwelle zu einer neuen Ära der AI stehen. Wir bewegen uns von reinen Sprachmodellen hin zu Systemen, die in der Lage sind, komplexe kognitive Aufgaben zu bewältigen. Die Zukunft der AI verspricht, unsere Vorstellungen von maschineller Intelligenz grundlegend zu verändern und neue Möglichkeiten in Wissenschaft, Technologie und darüber hinaus zu eröffnen.## Die Revolution der Spracherkennung

In den letzten Jahren haben wir eine bemerkenswerte Entwicklung im Bereich der Spracherkennung erlebt. Stellen Sie sich vor: Ihr kleines Handy, ausgestattet mit einem leistungsstarken Chip, ist heute in der Lage, aus Ihrem vielfältigen Wortschatz - inklusive Dialektfärbungen - das richtige Wort zu erkennen und es grammatikalisch korrekt in einen Satz einzufügen. Die Präzision dieser Technologie ist atemberaubend. Ohne zusätzliche Herausforderungen wie Umgebungsgeräusche liegt die Fehlerrate bei weniger als 2% - ein wahrhaft beeindruckendes Ergebnis.

### 1.7.2 Von Worten zu Bedeutungen

Doch die Entwicklung geht weit über die bloße Worterkennung hinaus. In den vergangenen zehn Jahren haben wir eine semantische Revolution miterlebt. Es geht nicht mehr nur darum, gesprochene Sätze korrekt zu transkribieren. Der eigentliche Durchbruch liegt in der Fähigkeit, die Bedeutung hinter den Worten zu erfassen. Moderne Systeme können nicht nur Sätze erkennen, sondern auch die darin enthaltenen Aussagen, Instruktionen oder Befehle verstehen und entsprechend darauf reagieren. Diese semantische Ebene eröffnet völlig neue Möglichkeiten der Mensch-Maschine-Interaktion.

## **1.8 Echtzeit-Übersetzung und ihre Implikationen**

### **1.8.1 Die nahe Zukunft der Sprachkommunikation**

Lassen Sie mich ein faszinierendes Szenario skizzieren, das in naher Zukunft Realität werden könnte: Stellen Sie sich vor, Sie haben auf Ihrem Smartphone eine App installiert, die mit einem Ohrhörer verbunden ist. Mit dieser Technologie könnten Sie eine Konversation führen, bei der Ihr Gegenüber in einer beliebigen Sprache spricht, und Sie hören die Übersetzung in Echtzeit in Ihrer eigenen Sprache. Noch faszinierender: Wenn Sie in Ihrer Muttersprache antworten, hört Ihr Gesprächspartner Ihre Worte unmittelbar in seiner Sprache.

Dies mag nach Science-Fiction klingen, doch tatsächlich existieren bereits alle notwendigen Komponenten. Es ist nur noch eine Frage der technischen Umsetzung und Integration.

### **1.8.2 Ein Gedankenexperiment zur Sprachübersetzung**

Lassen Sie uns dieses Konzept in einem Gedankenexperiment weiter ausführen. Stellen Sie sich vor, ich würde diese Vorlesung in einer Sprache halten, die keiner von Ihnen versteht - einer fiktiven Sprache, die nur ich beherrsche. Sie alle tragen Kopfhörer und hören meine Worte in Echtzeit übersetzt, sogar mit meinem charakteristischen Sprechstil.

Noch faszinierender: Die moderne Technologie ermöglicht es sogar, den Sprechstil anzupassen. Wenn Ihnen meine Art zu sprechen nicht zusagt und Sie beispielsweise die Ausdrucksweise von Jürgen Habermas bevorzugen, könnten Sie theoretisch die Übersetzung in seinem Stil hören. So würde Habermas' Stimme meine Vorlesung vortragen, obwohl ich hier unten stehe.

Dies mag zunächst absurd klingen, aber ich hoffe, dass wir noch in diesem Semester die Möglichkeit haben werden, ein solches Experiment durchzuführen. Ich arbeite ernsthaft daran, dies zu realisieren, und bis Februar sollte es machbar sein.

## **1.9 Die Grenzen der Sprache und des Verstehens**

### **1.9.1 Ausgestorbene Sprachen und künstliche Kommunikation**

Nun möchte ich Sie zu einem weiteren Gedankenexperiment einladen. Nehmen wir an, wir verwenden nicht eine beliebige moderne Sprache, sondern eine ausgestorbene Sprache, die heute niemand mehr als

Muttersprache spricht. Stellen Sie sich vor, ein Schauspieler oder Linguist trägt diese Vorlesung in einer solchen Sprache vor. Wie würden Sie diese Situation beurteilen?

Gehen wir noch einen Schritt weiter: Der Vortragende versteht weder Philosophie noch den Inhalt meiner Vorlesung. Er hat lediglich ein Manuskript erhalten, das er in der ausgestorbenen Sprache vorliest. Das Übersetzungsprogramm überträgt seine Worte in Ihre gewählte Sprache, vielleicht sogar im Sprechstil eines Kollegen Ihrer Wahl. Sie hören den Vortrag und gehen nach Hause mit dem Gefühl, einer ganz normalen, wenn auch etwas ungewöhnlich inszenierten Vorlesung beigewohnt zu haben.

### **1.9.2 Die Grenze zwischen Mensch und Tier in der Kommunikation**

Nun kommen wir zum letzten und vielleicht provokantesten Schritt unseres Gedankenexperiments. Stellen Sie sich vor, wir nehmen die akustischen Signale nicht von einem Menschen auf, sondern von einem Vogel - sagen wir, einer Nachtigall. Der Gesang des Vogels wird aufgenommen, verarbeitet und plötzlich hören Sie in Ihrem Ohrhörer die Worte "Ich möchte jetzt trinken" in der Stimme von Habermas.

Wie bewerten Sie diese Situation? Handelt es sich um Sprache? Um einen echten Inhalt? Oder ist es lediglich eine trainierte Lauterzeugung, ähnlich einem Papagei? Ich vermute, Ihre erste Reaktion wäre Skepsis. Sie würden sagen, dass es sich nicht um echte Sprache handelt, sondern nur um Laute, die wie Sprache klingen.

Doch was geschieht, wenn wir beginnen, eine Konversation zu führen? Ein menschlicher Zuhörer spricht in sein Mikrofon, und seine Worte werden in Nachtigallengesang umgesetzt. Plötzlich reagiert der Vogel darauf. Sie hören in Ihrem Kopfhörer: "Oh, hier gibt es kein Wasser. Schade." Und der Vogel fliegt weg.

Wie beurteilen Sie eine solche Interaktion? Es mag wie Science-Fiction klingen, aber solche Experimente werden bereits durchgeführt. Natürlich ist der Themenbereich bei Vögeln noch relativ eingeschränkt, hauptsächlich auf grundlegende Bedürfnisse wie Essen, Trinken und Gefahren beschränkt. Aber eine Art von Konversation findet statt.

Es gibt sogar Berichte über ähnliche Experimente mit Hunden, die angeblich sehr viel differenziertere Ausdrucks- und Themenzusammenhänge erörtern können. Beachten Sie, wie ich hier bereits Begriffe wie "erörtern" verwende - Begriffe, die wir traditionell nur der menschlichen Kommunikation zuschreiben würden.

## **1.10 Die philosophischen Implikationen**

### **1.10.1 Die Natur der Kommunikation**

Diese Szenarien werfen fundamentale Fragen auf: Was macht eine Konversation aus? Was bedeutet es, Gedanken auszutauschen? Wir sind es gewohnt zu denken, dass echte Kommunikation zwangsläufig menschliche Akteure erfordert. "Ich kann doch nicht mit meiner Wand reden", würden wir sagen. Aber nun haben wir Arrangements, die diese Annahme in Frage stellen.

In den kommenden Wochen werden wir wahrscheinlich mit Videos und Berichten über solche Phänomene konfrontiert werden. Wie sollen wir darauf reagieren? Es werden sicherlich hitzige Debatten entstehen. Manche werden Manipulation oder Täuschung vermuten. Aber diese Experimente sind real.

### **1.10.2 Die Verschiebung unserer Wahrnehmung**

Bemerken Sie, wie ich in meiner Beschreibung dieser Szenarien allmählich dazu übergegangen bin, von "Konversation", "Mitteilung" und "Gedankenaustausch" zu sprechen - und zwar im Kontext der Interaktion zwischen Vögeln und Menschen? Ich hoffe, Sie haben diesen sprachlichen Übergang nicht als besonders befremdlich oder absurd empfunden.

Mein Ziel war es zu zeigen, wie bereit wir sind, aufgrund der Komplexität der beobachteten Prozesse, eine Situation als echte Kommunikation zu interpretieren. Wenn wir erkennen, dass Erwartungen ausgedrückt und erfüllt werden, dass die Regeln sprachlicher Konversation eingehalten werden, neigen wir dazu, die biologische Natur des Akteurs als zweitrangig zu betrachten.

### **1.10.3 Philosophische Fragen zur Natur des Verstehens**

Diese Überlegungen führen uns zu grundlegenden philosophischen Fragen:

1. Wie können wir den Inhalt einer Kommunikation überprüfen?
2. Was rechtfertigt unseren Übergang von der Beurteilung eines reinen Lautes ohne Sinn zu einer bedeutungsvollen Äußerung?
3. Können wir wirklich zwischen der materiellen Konstitution einer Äußerung (ein Audiosignal, ein geschriebener Buchstabe) und einem begleitenden Gedanken unterscheiden?

4. Ist das Vorhandensein eines “Gedankens” notwendig, um von echtem Inhalt und echter Konversation zu sprechen?

Diese Fragen berühren den Kern der Sprachphilosophie und des Verstehens. Sie erinnern an Wittgensteins Überlegungen zur Natur der Sprache und des Verstehens.

## **1.11 Empirische Perspektiven auf tierische Kommunikation**

### **1.11.1 Reiz-Reaktions-Mechanismen vs. Bedeutung**

Ein wichtiger Aspekt in dieser Diskussion ist die Frage, ob tierische Kommunikation lediglich auf Reiz-Reaktions-Mechanismen beruht oder ob es eine tiefere Ebene des Verstehens gibt. Tiere haben zweifellos biologisch implementierte Mechanismen, die es ihnen ermöglichen, auf Reize zu reagieren und entsprechende Signale auszusenden. Diese Prozesse sind physiologisch messbar und können oft ohne Rückgriff auf Konzepte wie “Bedeutung” oder andere geistige Kompetenzen erklärt werden.

Doch reicht diese Erklärung aus? Können wir wirklich alle Aspekte tierischer Kommunikation auf rein physiologische Prozesse reduzieren? Oder gibt es Anzeichen für ein tieferes Verständnis, das über simple Reiz-Reaktions-Muster hinausgeht?

### **1.11.2 Die Rolle der Technologie in der Interpretation tierischer Kommunikation**

Die Entwicklung von “Tier-Übersetzungs-Apps” fügt dieser Diskussion eine weitere Dimension hinzu. Diese Apps nehmen die Signale von Tieren auf - sei es das Schwanzwedeln eines Hundes oder der Gesang eines Vogels - und übersetzen sie in für Menschen verständliche Sätze. Obwohl dies technisch machbar ist, wirft es wichtige Fragen auf:

1. Wie genau können solche Übersetzungen sein?
2. Interpretieren wir möglicherweise zu viel in die tierischen Signale hinein?
3. Wo liegt die Grenze zwischen echter Kommunikation und unserer Projektion von Bedeutung?

Diese Fragen führen uns zurück zu den grundlegenden philosophischen Problemen des Verstehens und der Bedeutung. Sie zwingen uns, unsere Vorstellungen von Sprache, Kommunikation und sogar Bewusstsein zu überdenken.## Einführung in die Mensch-Tier-Kommunikation

In unserer heutigen Vorlesung befassen wir uns mit einem faszinierenden Thema: der Kommunikation zwischen Mensch und Tier. Wir werden uns insbesondere auf ein Beispiel konzentrieren, das die Komplexität und die Herausforderungen dieses Forschungsgebiets verdeutlicht.

### **1.11.3 Der sprechende Papagei - Ein Fallbeispiel**

Stellen Sie sich vor, wir hätten ein technisches Modell entwickelt, das in der Lage ist, die Laute eines Papageien zu interpretieren. Dieses Modell, empirisch validiert und sorgfältig geprüft, hat eine klare Zuordnung getroffen: Wenn der Papagei einen bestimmten Laut von sich gibt - nennen wir ihn "Piep" mit drei "E" - dann bedeutet das "Wasser".

Doch was bedeutet diese Zuordnung wirklich? Können wir tatsächlich davon ausgehen, dass der Papagei bewusst kommuniziert? Oder handelt es sich lediglich um eine Imitation ohne tiefere Bedeutung?

### **1.11.4 Herausforderungen der Interpretation**

Bei der Interpretation tierischer Kommunikation stoßen wir auf mehrere Herausforderungen:

1. Intentionalität: Wie können wir sicher sein, dass der Papagei tatsächlich die Absicht hat, uns etwas mitzuteilen?
2. Kontextabhängigkeit: Inwieweit spielt der Kontext eine Rolle bei der Interpretation der Laute?
3. Komplexität der Kommunikation: Wie komplex sind die Absichten und Bedürfnisse, die ein Tier ausdrücken kann?
4. Übersetzungsproblematik: Ist es angemessen, tierische Laute in menschliche Sprache zu übersetzen?

### **1.11.5 Die Frage der Bedeutung**

Ein wesentlicher Punkt in unserer Diskussion ist die Frage nach der Bedeutung. Wenn wir dem Papagei zugestehen, dass er "Wasser" kommunizieren kann, implizieren wir damit ein gewisses Maß an kognitivem Vermögen. Doch wie weit reicht dieses Vermögen? Kann ein Papagei wirklich ein Konzept wie "Durst" haben, oder projizieren wir hier menschliche Konzepte auf tierisches Verhalten?

### **1.11.6 Vergleich zur KI-Debatte**

Interessanterweise ähnelt diese Diskussion in vielerlei Hinsicht der Debatte um künstliche Intelligenz. Während wir bei Tieren oft bereit sind, komplexe kognitive Fähigkeiten anzunehmen, sind wir bei Maschinen oft skeptischer. Dies wirft die Frage auf: Welche Kriterien legen wir an, um "echte" Kommunikation oder Intelligenz zu definieren?

### **1.11.7 Ethische und philosophische Implikationen**

Die Frage, ob und wie wir mit Tieren kommunizieren können, hat weitreichende ethische und philosophische Implikationen:

- Wenn Tiere komplexe Gedanken und Gefühle ausdrücken können, wie verändert das unsere moralischen Verpflichtungen ihnen gegenüber?
- Welche Rolle spielt Sprache in unserem Verständnis von Bewusstsein und Intelligenz?
- Inwiefern unterscheidet sich tierische Kommunikation von menschlicher Sprache, und was sagt das über die Natur von Bedeutung und Verständnis aus?

## 2 Was ist AI?

### 2.1 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 Verwandtschaft 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]

## 2.2 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.



**Angebot**

**WW8400D, BESPOKE Waschmaschine, QuickDrive, AI Wash, 1-11 kg, EEK: A (-40%)**

WW11DB8B95GHU2

☆☆☆☆☆ 0 (0)

**330 € Energieeinsparung**  
Die beste Wahl zum Energiesparen

Im Vergleich zum ineffizientesten Modell dieser Art (DE)

- 40% mehr Energieersparnis als Effizienzklasse A
- AI Ecobubble™
- QuickDrive Option - viele Programme bis zu 50% schneller
- Automatische Waschmittel-Dosierung und AI Wash
- SuperSpeed 39 Min.
- Klartext-Display mit AI Control

**Figure 2.1:** 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.

### 2.2.1 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.

### **2.2.2 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 Technologiesgeschichte 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.

### **2.2.3 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.

### **2.2.4 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 kuratiert waren. Doch in der nächsten Phase der digitalen Revolution wird sich dies grundlegend ändern.

### **2.2.5 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.

## **2.3 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.

### **2.3.1 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.

### **2.3.2 Simultanübersetzung und Lektoratsassistenz**

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.

### **2.3.3 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.

### **2.3.4 Das Labor *Lettre AI***

In eigenen Labor - *Lettre AI* (*Lettre* (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.

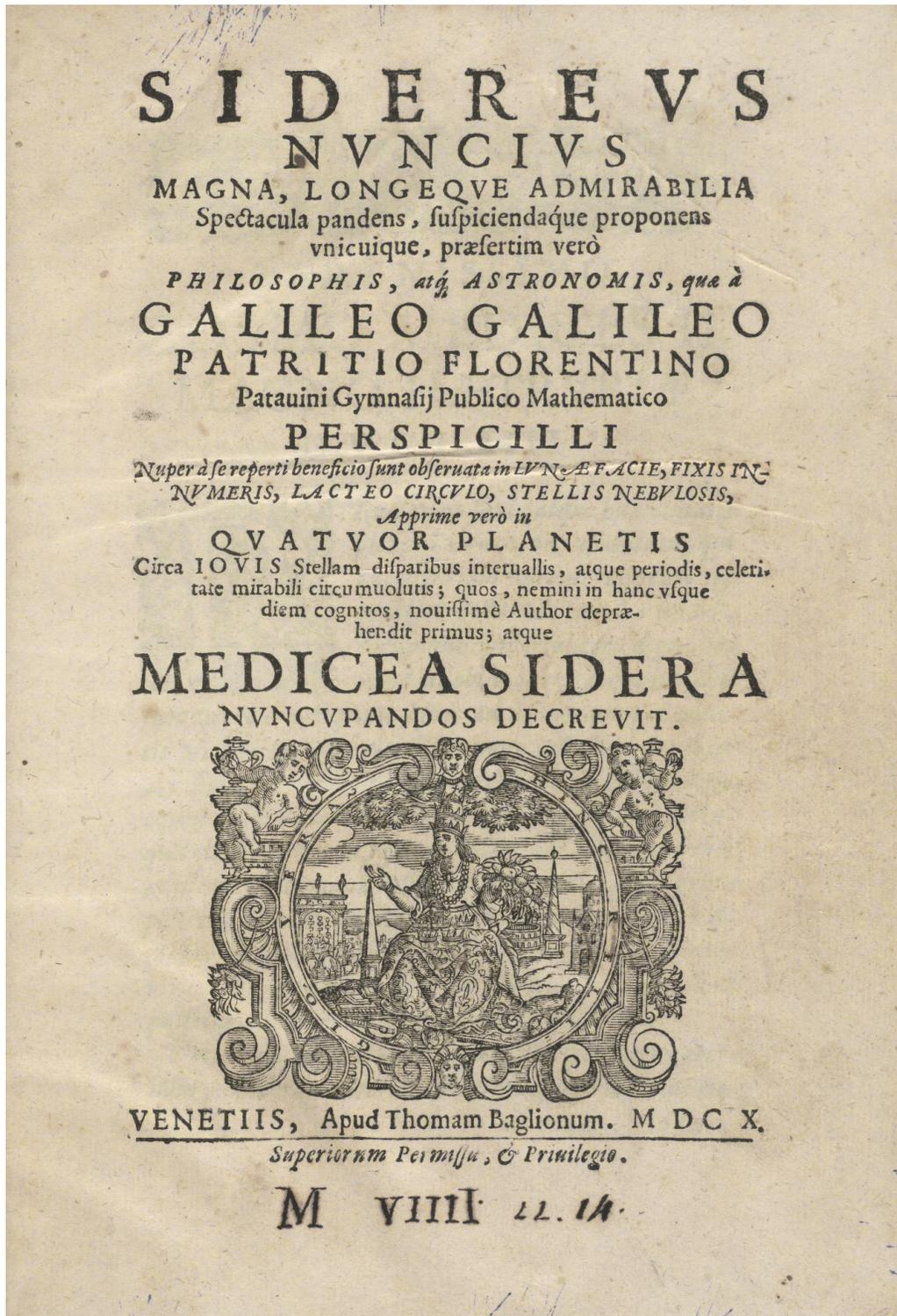


Figure 2.2: Sidereus Nuncius

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.

### **2.3.5 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.

### **2.3.6 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.

## **2.4 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.

### **2.4.1 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.

### **2.4.2 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.

## **2.5 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.

### **2.5.1 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.

### **2.5.2 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.

### **2.5.3 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 unerkant 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.

#### **2.5.4 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 Unhintergebarkeit 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.

#### **2.5.5 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.

### **2.5.6 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.

### **2.5.7 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.

### **2.5.8 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.

### **2.5.9 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.

### **2.5.10 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.

## **2.6 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.

## **2.7 Bisläng nicht lösbare 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.

### **2.7.1 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.

### **2.7.2 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.

### **2.7.3 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, welche enormen 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.

### **2.7.4 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.

## 2.8 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 grammatikalisch 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.

### **2.8.1 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.

### **2.8.2 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.

### **2.8.3 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 Meisterinterpreten.

#### **2.8.4 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.

#### **2.8.5 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.

### **2.8.6 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 Gesamten 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.

## **3 Die Revolution der AI**

### **3.1 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.

### **3.2 Traditionell schwer lösbare 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.

#### **3.2.1 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.

### **3.2.2 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.

### **3.3 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 lösbare 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.

### **3.4 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.

#### **3.4.1 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.

### **3.4.2 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.

### **3.4.3 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.

## **3.5 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.

### **3.5.1 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.

### **3.5.2 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.

### **3.5.3 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.

### **3.5.4 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.

### **3.5.5 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.

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

Sie sehen, welche weitreichenden 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?

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

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

### **3.6.1 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.

#### **3.6.1.1 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.

### **3.6.2 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 dasselbe aus, haben denselben Wahrheitswert?

### **3.6.2.1 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 Grundausstattung 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.

### **3.6.2.2 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.

### **3.6.3 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.

## 3.7 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.

### 3.7.1 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.

### 3.7.2 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. Plöt-

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

### **3.7.3 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.

## **3.8 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.

### **3.8.1 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.

### **3.8.2 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.

## 3.9 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.

### 3.9.1 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.

### **3.9.2 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.

## **3.10 Erweiterung der AI-Modelle**

### **3.10.1 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.

### **3.10.2 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.

## **3.11 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.

### **3.11.1 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.

### **3.11.2 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.

## **3.12 Anwendungsbeispiele und Potenziale**

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

### **3.12.1 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.

### 3.12.2 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 Bedeutungsklä rung 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 Pommestube betreibt.

Durch den anschließenden Dialog, in dem Sie klarstellen, dass Sie nicht den Mathematiker, sondern den Pommestuben-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.

## **4 Charakter von LLMs**

### **4.1 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.

### **4.2 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.

### **4.2.1 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 Sprachverstehens und -verarbeitens ist fundamental für alle Modelle der künstlichen Intelligenz, mit denen wir es hier zu tun haben.

### **4.2.2 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.

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

### **4.3.1 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.

### **4.3.2 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.

#### **4.3.2.1 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)

#### **4.3.2.2 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.

#### **4.3.3 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.

#### **4.3.4 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.

#### **4.3.5 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.

##### **4.3.5.1 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.

#### **4.3.6 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.

### **4.3.7 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.

### **4.3.8 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.

#### 4.3.8.1 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.

#### 4.3.9 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.

#### **4.3.10 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.

### **4.4 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.

#### **4.4.1 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.

##### **4.4.1.1 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.

#### **4.4.2 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.

##### **4.4.2.1 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.

#### **4.4.3 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.

### **4.5 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.

#### **4.5.1 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.

## **4.6 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.

### **4.6.1 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.

### **4.6.2 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.

## 4.7 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 Sprachverstehens
- 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.

### 4.7.1 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 Kontextkonstrukts. 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.

#### **4.7.2 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.

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

#### **4.8.1 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.

#### **4.8.2 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 überschneiden 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.

### **4.8.3 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.

## **4.9 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.

# 5 LLM für Sprache

## 5.1 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.

## 5.2 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.

### **5.2.1 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.

## **5.3 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.

### **5.3.1 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.

## 5.4 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.

### 5.4.1 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.

### 5.4.2 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.

### 5.4.3 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.

## 5.5 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.

## 5.6 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.

### **5.6.1 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.

### **5.6.2 Weitere Lernmöglichkeiten**

Die Lernmöglichkeiten gehen jedoch noch viel weiter. Die Hersteller der Modelle bieten Ihnen beispielsweise 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.

### **5.6.3 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.

## **5.7 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.

## 5.8 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.

## 5.9 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?

### **5.9.1 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.

### **5.10 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.

## 5.11 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.

### **5.11.1 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.

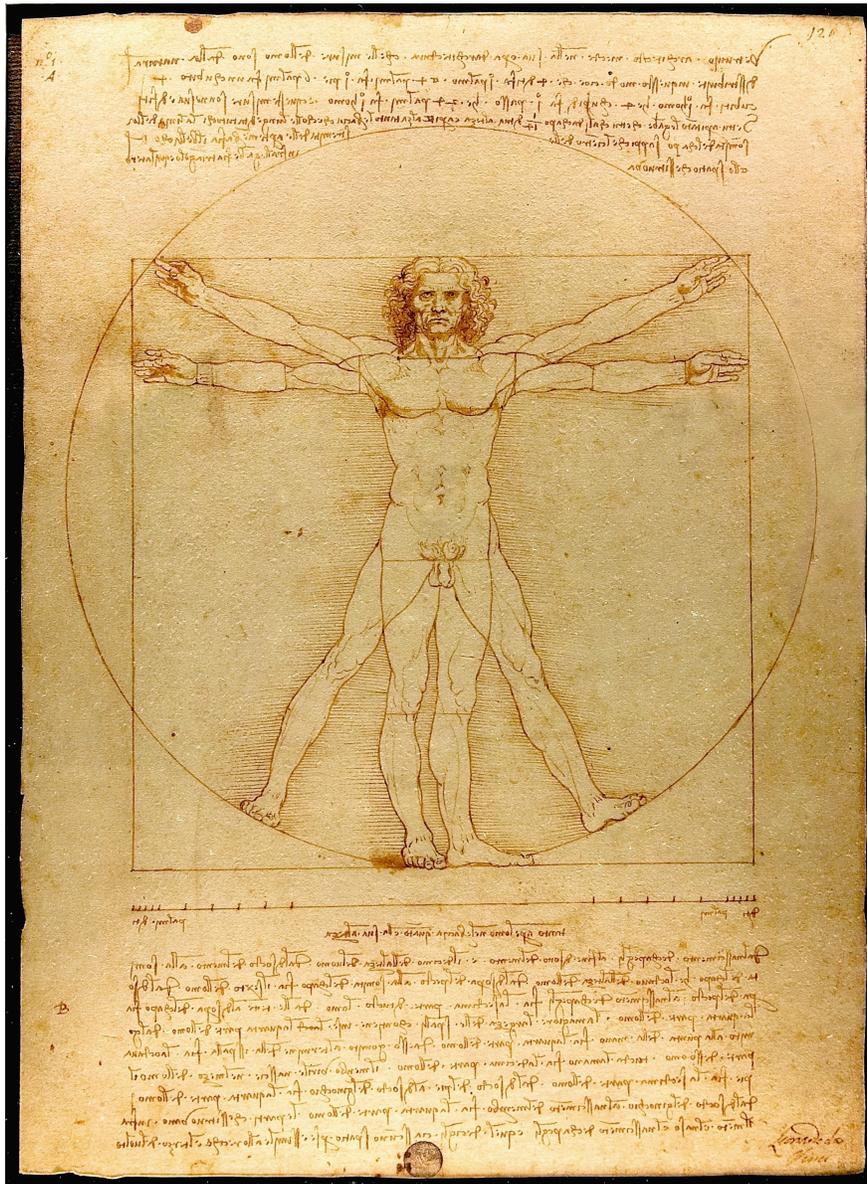
Ein solches AI-Modell folgt keinem vorgeschriebenen Verhalten, sondern agiert als Charakter, als zu schaffendes Individuum. Wir sollten es behandeln wie eine Entität, von der wir bestimmte Leistungen und Funktionen erwarten und mit der wir entsprechend agieren können.

## **5.12 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.

### **5.12.1 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 5.1:** Der vitruvische 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.

### **5.12.2 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 5.2:** 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.

### 5.12.3 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.



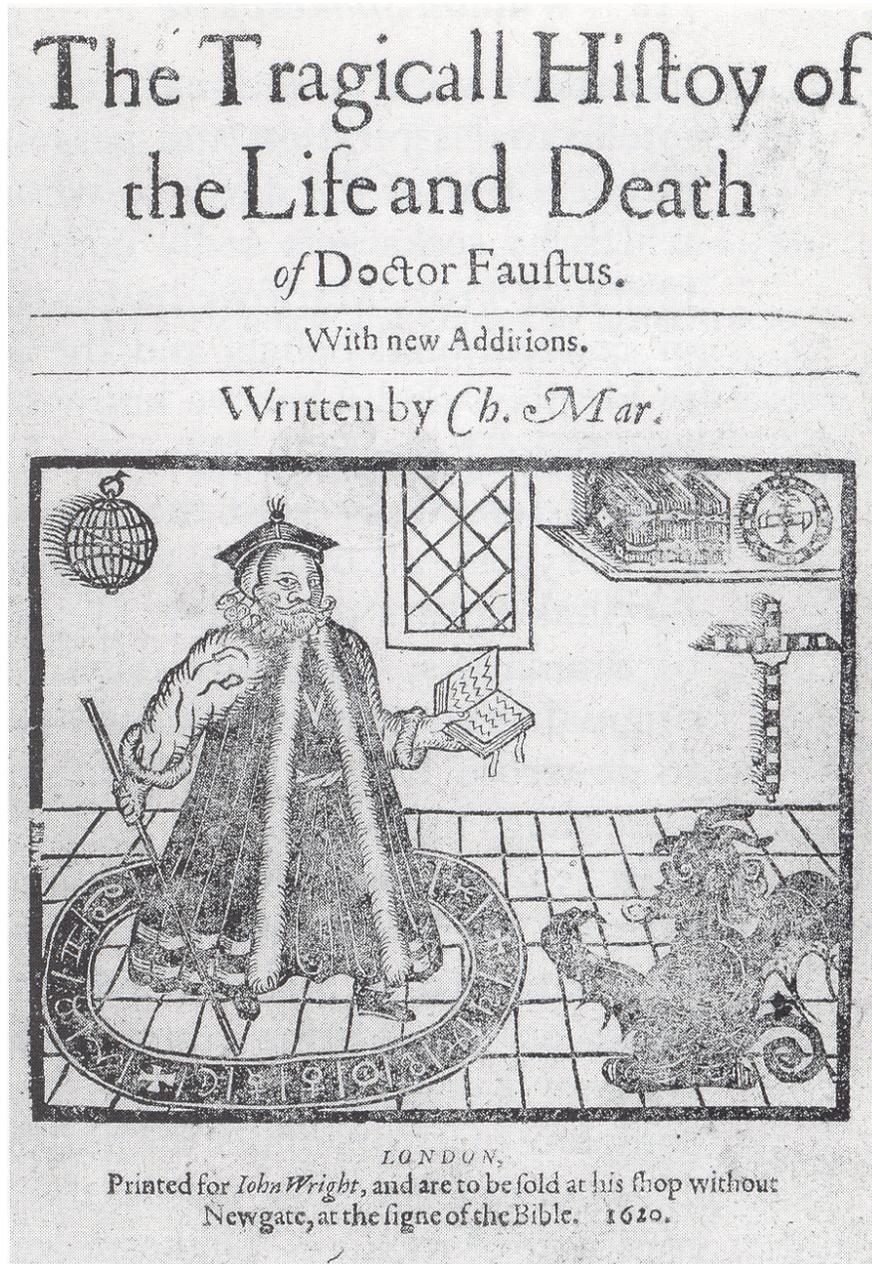
**Figure 5.3:** Beuys und der tote Hase

Übertragen auf AI-Modelle bedeutet dies, dass auch sie in der Lage sein müssen, ihre Ergebnisse und The- sen 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.

### **5.13 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 5.4:** 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 ergünden, 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?

### **5.13.1 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.

### **5.13.2 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.

### 5.13.3 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.

## 5.14 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.

### **5.14.1 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.

## **5.15 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.

## 5.16 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!

## 6 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.

### 6.1 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 zumindest 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.

#### 6.1.1 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 Manuskript 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.

## **6.2 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.

### **6.2.1 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.

## **6.3 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.

### **6.3.1 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.

## **6.4 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.

### **6.4.1 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.

Wohlgermerkt 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.

#### **6.4.2 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 erstarren: 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.

### **6.4.3 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.

### **6.4.4 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.

### **6.4.5 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.

## **6.5 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.

## 6.6 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.

### 6.6.1 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 Ausgabertexte 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.

## 6.7 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.

## 6.8 Charakteristika und Fähigkeiten der Modelle

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

### 6.8.1 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.

### 6.8.2 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.

### **6.8.3 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 grammatikalische Flexionen wie Deklinationen berücksichtigt.

### **6.8.4 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.

## **6.9 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.

### **6.9.1 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.

### **6.9.2 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.

### **6.9.3 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.

## **6.10 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.

### **6.10.1 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.

## **6.11 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.

### 6.11.1 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.

### 6.11.2 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 seltenere 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 lösbar 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!

# 7 Denken mit Logik

## 7.1 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.

### 7.1.1 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.

### **7.1.2 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.

### **7.1.3 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 Aufgabe 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 Textverarbeitungscompetenz, auf der dieses Projekt aufbauen soll.

## **7.2 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.

### **7.2.1 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.

### **7.2.2 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, schlichtweg. 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.

### 7.2.3 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>

#### 7.2.3.1 Context of Discovery

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

#### 7.2.3.2 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 heißt 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 heißt, postfaktum der Entdeckung wird über das Ergebnis publiziert, und zwar in rechtfertigender Weise.

### 7.2.4 Mangel an Discovery-Prozessen in Publikationen

Sie finden kaum, ich möchte fast wagen zu sagen, im Unterpromillebereich, wissenschaftliche Publikationen - und ich habe das mal probenhalber 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?

#### **7.2.4.1 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.

#### **7.2.4.2 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.

## 7.2.5 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.

### 7.2.5.1 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.

### **7.2.5.2 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 kritischen, 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.

## **7.3 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.

### **7.3.1 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.

## **7.4 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.

### **7.4.1 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.

#### **7.4.2 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.

### **7.5 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.

#### **7.5.1 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 interessanterweise 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.

### 7.5.2 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.

### 7.5.3 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 Inkompetenz 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.

#### **7.5.4 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.

### **7.5.5 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?

### **7.5.6 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 stündlich, basierend auf dem Feedback der Community. Allerdings bleibt die grundlegende Sprachkompetenz der Modelle unverändert, da eine Aktualisierung dieser enormen Aufwand und Kosten bedeuten würde.

### **7.5.7 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.

## 7.6 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.

### 7.6.1 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.

### 7.6.2 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.

### 7.6.3 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.

## 7.7 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.

### **7.7.1 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.

## **7.8 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 Programmmodule 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.

## 7.9 Die Bedeutung der Schlüssigkeit

Die Schlüssigkeit von Aussagen und Argumenten ist von eminenter 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.

### 7.9.1 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.

## 7.10 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.

### 7.10.1 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 Konstrukts von Sätzen zu beurteilen.

## 7.11 Ein praktisches Beispiel

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

1. Wenn die Menschheit zu viel CO2 erzeugt, steigt der Wasserspiegel des Ozeans.
2. Der Lebensstandard in Italien ist sehr hoch und die Menschheit erzeugt zu viel CO2.
3. Der Lebensstandard in Indien ist nicht so hoch wie in Italien.

Konsequenz: Also steigt der Wasserspiegel.

### 7.11.1 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>

### 7.11.2 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 epikureische 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.

## 7.12 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 Wahrheitstafeln
- Formalisierung der logischen Beziehung der einzelnen Sätze
- Einsatz der Wahrheitstafel als systematisches Instrument
- Überprüfung der Validität des Verfahrens

Das Ergebnis: Das Argument ist schlüssig.

### **7.13 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.

# 8 Argumente

## 8.1 Begrüßung und Einleitung

Guten Tag, meine Damen und Herren. Herzlich willkommen zur siebten Vorlesung unseres Kurses - wir haben also bereits Halbzeit erreicht. Heute möchte ich Ihnen einen spezifischen Bereich vorstellen, der die Philosophie im Kontext der KI-Funktionsweise besonders herausfordert. In den bisherigen Modellen und Technologien sind philosophische Konzepte bisher nur rudimentär umgesetzt. Ich werde Ihnen zeigen, an welchen Stellen dies der Fall ist und welchen Nutzen philosophische Beiträge und Analysen für eine angemessene Funktionsweise der KI haben können.

### 8.1.1 Rasante Entwicklung der KI-Technologie

Einige von Ihnen haben in der Vorlesung die Beispiele parallel mit den entsprechenden Chatangeboten ausprobiert und dabei festgestellt, dass sich die Ergebnisse und Funktionsweisen der jeweiligen Modelle täglich ändern. Lassen Sie sich nicht irritieren, wenn die Resultate, die Sie hier in der Vorlesung sehen, am Abend in Ihrer eigenen technologischen Umgebung anders ausfallen. Das gilt auch für die heutigen Beispiele.

Ich werde erneut auf das logische oder aussagenlogische Schließen eingehen. Bei den von mir genutzten Modellen beobachte ich, dass die Befundlage bezüglich der Schlüssigkeit eines Arguments von einem Tag auf den anderen schwankt. Wir werden gleich sehen, warum das so ist - zufriedenstellend ist es jedenfalls nicht.

Die KI-Technologie entwickelt sich derzeit in atemberaubendem Tempo. Wenn Sie Zeit haben, empfehle ich Ihnen, sich nächsten Montag die Präsentation von Apple anzuschauen. Das Unternehmen wird seine neue KI-Maschine namens Ajax vorstellen, die in jede Apple-Umgebung integriert sein wird - vom Notebook bis zur neuesten iPad-Generation.

Ich hatte bereits die Gelegenheit, eine Vorabversion zu testen. Zwar gibt es hier und da noch ein paar Kinderkrankheiten, aber das ist bei ersten Versionen nicht ungewöhnlich. Der Einsatz und die Funktion-

alität der KI werden sich in Zukunft radikal verändern und alle alltäglichen Anwendungen der Informatik durchdringen.

Als besondere Herausforderung möchte ich Ihnen in der nächsten Vorlesungswoche eine dieser Anwendungen zeigen. Ich will nur eine kryptische Andeutung machen: Wenn alles funktioniert, wird in alle Apple-Geräte eine Augenbewegungsdetektion integriert sein. Damit lässt sich erkennen oder abfragen, was Sie gerade auf dem Bildschirm anschauen. Das kann so weit zur Steuerung genutzt werden, dass eine Art Gedankenlesen stattfindet. Sie müssen dann keinen Text mehr mit der Maus auswählen, um etwas damit zu machen, sondern starren die gewünschte Passage einfach an. Anschließend fixieren Sie einen Ausführungsknopf mit Ihrem Blick - und schon wird die entsprechende Aktion ausgelöst. Ob und wie genau das funktioniert, werden wir uns anschauen. Hoffentlich verlieren wir uns dabei nicht zu sehr in Science-Fiction, sondern erahnen zum einen das rasante Entwicklungstempo und zum anderen die Mittel, mit denen diese Technologien zusammengeführt und integriert werden.

Der besondere technologische, ökonomische und anwendungsbezogene Reiz liegt nämlich gerade in der Verknüpfung der verschiedenen Elemente. Es geht nicht nur um die Funktionalität der KI selbst, sondern um deren nahtlose Einbindung in eine generelle Arbeits- und Nutzungsumgebung. Das reicht bis hin zum sogenannten Smart Home, bei dem diverse Geräte in einer Wohnung über eine solche Arbeitsumgebung ferngesteuert werden können. Doch dazu vielleicht mehr in der nächsten Stunde.

## **8.2 Sprachverarbeitung durch KI**

Kommen wir zunächst noch einmal zum basalen Kern der KI-Funktionalität in Bezug auf Sprache zurück. Wir hatten ein beispielhaftes Textschnipsel thematisiert, das stellvertretend für alle Texte beliebiger Größe und Herkunft steht.

### **8.2.1 Allgemeingültigkeit des Beispiels**

Was wir hier haben, ist nicht einmal ein bewusst von mir konstruiertes Zitat, sondern ein Satzfragment. Doch solche Textfragmente sind charakteristisch für alles, womit wir im Umgang mit Texten konfrontiert sind - seien es Schnipsel, Ausschnitte oder eine Art Zeitungsschnipsel. Von diesen wollen wir etwas wissen und mit ihnen arbeiten. Wie wir das mithilfe der KI tun können und welche philosophischen Konzepte hinter jeder KI stehen müssen, werde ich Ihnen heute verdeutlichen.

Die Befragung und Verarbeitung solcher Texte ist zunächst sehr generell. Der hier gezeigte Schnipsel steht exemplarisch für Textausschnitte jeglicher Größe und Provenienz.

Das gilt auch für das Goethe-Projekt, bei dem es um Dokumente geht, die Goethes Biografie betreffen. Wir haben dabei drei Hauptgruppen: Goethes eigene Schriften, seinen Briefwechsel und Dokumente, die sich auf sein Leben beziehen. Das umfasst alles, woran Goethe selbst beteiligt war oder was ihn unmittelbar betraf.

### **8.2.2 Große Dokumentenbestände**

Doch das ist bei weitem nicht alles an Dokumenten, die in einem solchen Fall berücksichtigt werden müssen. Denn dazu gehört auch jegliches Material, das für den historischen Kontext relevant ist, um beliebige Anfragen an kleinste Schnipsel dieser Art sinnvoll beantworten zu können. Das schließt Dokumente von Zeitgenossen, historische, ökonomische und politische Schriftstücke mit ein.

In der heutigen Zeit der Nachdigitalisierung von Archiven und Bibliotheken liegt der Großteil dieser Dokumente bereits in digitaler Form vor, meist als PDF. Für historische Persönlichkeiten wie Goethe sind sie häufig sogar online zugänglich. Das Material ist also vorhanden.

### **8.2.3 Begrenzte Nutzbarkeit derzeitiger Digitalisate**

Doch was können wir heute mit diesem Material anfangen? Im besten Fall lassen sich die PDF-Dokumente lesen - darüber hinaus ist man auf den eigenen Intellekt und das persönliche Textverständnis angewiesen. Das ist der Normalfall.

Was jetzt für die geisteswissenschaftliche und kulturhistorische Forschung ansteht, ist die inhaltliche Erschließung dieses Gesamtkorpus. Und genau diese Erschließung der schriftlichen Überlieferung wird sich der Methoden bedienen, die wir hier als KI für wissenschaftliche Inhalte thematisieren.

Derzeit sind diese Inhalte für die gängigen KI-Modelle weder zugänglich noch werden sie von ihnen verwertet. Auch auf die Archive wird kein Bezug genommen, sie werden nicht ausgewertet. Man kann keine sinnvollen Abfragen stellen, um beispielsweise einen Chat zu starten und nach einer bestimmten Aussage Goethes zu suchen. Auf solche Fragen wird man keine brauchbaren Antworten erhalten.

### **8.2.4 Philosophische Konzepte hinter der KI-Sprachverarbeitung**

Doch warum ist das so? Um das zu verstehen, befassen wir uns zunächst mit dem Inhalt und den Problemen unseres beispielhaften Textschnipsels. Es handelt sich um einen oder mehrere Sätze, einen kleinen, künstlich zusammengeschriebenen Textausschnitt. Er lautet wie folgt:

“Wenn über Jahre durch die Menschheit zu viel CO2 erzeugt wird, steigt der Wasserspiegel des Ozeans. Der Lebensstandard in Italien ist sehr hoch und die Menschheit erzeugt zu viel CO2. Also steigt der Wasserspiegel.”

Das mag unscheinbar wirken, doch philosophisch hat es dieser Text durchaus in sich. Er ist zwar nicht schwer zu verstehen, aber auch nicht trivial.

### **8.2.5 Verarbeitung durch KI-Modelle**

Wenn wir nun einen solchen Textschnipsel an ein GPT oder eine andere KI schicken, wissen Sie mittlerweile, was passiert: nämlich alles Mögliche. Denn das Programm bekommt zunächst keine Vorgabe, was es mit dem Schnipsel anfangen soll.

Das simulieren wir natürlich für die Vorlesung direkt. Ich habe hier einmal die zwei Hauptvertreter von Chat-Modellen genommen und den Textblock hineinkopiert. Die Chats sind so eingestellt, dass sie diesen Input als Aufforderung verstehen, etwas damit zu tun - ohne dass das explizit gesagt wird.

Sie können sehen, dass sich die Reaktionen von Maschine zu Maschine und von Tag zu Tag unterscheiden. Es gibt radikale Verweigerer, die höflich fragen, was Sie eigentlich von ihnen wollen. Aber die meisten Modelle fangen gleich an zu spekulieren, nach dem Motto: Sie wirken besonders intelligent, wenn Sie erst einmal Ihre eigenen Interessen einfüllen.

### **8.2.6 Verhalten von Chat-GPT**

Genau das macht Chat-GPT hier. Die Voreinstellung lautet: Identifiziere ein Thema und liefere dann eine Standarderklärung dazu, wie sie etwa in Wikipedia zu finden wäre. Das ist die Grundhaltung, ohne dass man auch nur den Hauch eines Hinweises gegeben hätte.

GPT sagt dann beispielsweise etwas über die Auswirkungen von Kohlendioxid-Emissionen auf den Meeresspiegel. Das ist eine heikle Frage, die ich bewusst gestellt habe, um herauszufinden, wie das System zur Debatte um Klimafolgen und Klimawandelleugner steht.

Hier wird es interessant: Einige KI-Maschinen haben eingebaute Sensoren und schneiden gerade beim Thema Klimaleugner gewisse Informationen ab - selbst wenn man sie aus rein wissenschaftlich-politischem Interesse abfragt. Etwa, welche Politiker im Bundestag sich jemals kritisch zum Klimawandel geäußert haben.

Die derzeitige Voreinstellung der Unternehmen hinter den Modellen, in diesem Fall OpenAI und Anthropic, ist es, mehr oder weniger stark in die Antworten einzugreifen. Das hatten wir bereits in der ersten Stunde thematisiert: Eine erhebliche Schwierigkeit und Gefahr dieses Bereichs besteht darin, dass eine solche Datenselektion - man muss es nicht Zensur nennen - durch die Modelle stattfindet und wohl auch unvermeidbar ist.

Es müssen Wege gefunden werden, diese Selektion selbst mitzubestimmen, was momentan nicht der Fall ist. Ich möchte nur darauf hinweisen, dass selbst bei scheinbar harmlosen Fragen im Hintergrund potenziell ein Zensurfilter am Werk ist, der bestimmte Informationen unterdrückt.

### **8.2.7 Nutzerseitige Steuerungsmöglichkeiten**

Wir mögen es positiv sehen, dass völlig absurde Vorstellungen und Thesen, wie sie in den sozialen Medien kursieren, bei den Antworten nicht mehr berücksichtigt werden. Doch wir haben keinerlei Einblick in die zugrundeliegenden Mechanismen und Algorithmen - und noch weniger die Möglichkeit, hier selbst steuernd einzugreifen. Unsere diesbezügliche Autonomie ist derzeit nicht vorhanden. Das ist zunächst keine Wertung, sondern eine Beschreibung der Gegebenheiten.

## **8.3 Erste Befunde KI-Modelle**

Lassen Sie uns nun das gleiche Experiment mit einem anderen Modell durchführen, um zu sehen, wie unterschiedlich die Maschinen mit dem Input umgehen. Wir probieren es mit Claude von Anthropic, der etwas anders eingestellt ist.

OpenAIs Chat-GPT agiert wie eine universelle Frage-Antwort-Maschine à la Google: Sie stellen eine Frage und das System versucht, diese maximal umfassend zu beantworten, indem es immer mehr standardisierte Referenzen einbezieht. Zukünftig sollen auch die verwendeten Quellen angegeben werden - ein Punkt, der derzeit noch häufig kritisiert wird. Denn momentan enthalten solche sachlichen Antworten keinerlei Quellenangaben, sodass man ihre Herkunft weder prüfen noch einschätzen kann.

Dieses Problem ist seit langem bekannt und einige Maschinen geben die Quellen bereits jetzt aus. Microsofts Chat-GPT macht das in einer Zusatzkomponente namens Chat-Pilot, und dieser Ansatz wird sich weiter verbreiten. Die hier gezeigte Antwort entspricht also noch dem Stand von vorgestern - morgen wird sie bereits anders aussehen.## Philosophische Grundlagen für die Weiterentwicklung von AI-Modellen

In den letzten Jahren ist das Interesse an der Philosophie im Kontext der Künstlichen Intelligenz (AI) merklich gewachsen. Die theoretische Philosophie hat sich im vergangenen Jahrhundert intensiv mit Fragestellungen beschäftigt, die heute für die Weiterentwicklung der AI von höchster Relevanz sind, bisher aber noch nicht ausreichend genutzt werden.

### **8.3.1 Grenzen aktueller AI-Modelle bei der Informationsverarbeitung**

Aktuelle AI-Modelle wie ChatGPT oder Claude von Anthropic zeigen bereits beeindruckende Fähigkeiten, wenn es darum geht, auf Nutzereingaben einzugehen und themenbezogene Antworten zu generieren. Allerdings stoßen sie schnell an ihre Grenzen, wenn es um die Bewertung der Gültigkeit, Wichtigkeit und Seriosität der verwendeten Quellen geht. Eine fundierte Abwägung der zugrunde liegenden Informationen findet nicht statt.

### **8.3.2 Philosophische Ansätze zur Verbesserung von AI-Modellen**

Die Philosophie bietet hier wertvolle Ansatzpunkte, um die Leistungsfähigkeit von AI-Modellen zu verbessern. Ein Kernthema ist dabei die Identifikation und Beurteilung von Argumenten. AI-Modelle müssen in der Lage sein, Argumente in Texten zu erkennen, deren Struktur zu analysieren und ihre Schlüssigkeit zu bewerten. Dies ist eine zentrale Fähigkeit für jede Form von AI, die den Namen "Intelligenz" verdient.

#### **8.3.2.1 Rekonstruktion von Argumenten als Herausforderung**

Die korrekte Rekonstruktion von Argumenten stellt AI-Modelle vor große Herausforderungen. Selbst scheinbar irrelevante Informationen können im Gesamtkontext eines Arguments plötzlich an Bedeutung gewinnen. Eine vorschnelle Aussortierung vermeintlich irrelevanter Details ist daher problematisch. Stattdessen müssen AI-Modelle lernen, potenzielle kausale Zusammenhänge zwischen verschiedenen Faktoren zu erkennen und zu bewerten.

#### **8.3.2.2 Trainingsmodelle für kausales Schließen**

Derzeit fehlt es an geeigneten Trainingsmodellen, um AI-Modelle im Bereich des kausalen Schließens zu schulen. Die Modelle füllen diese Lücke oft durch Plausibilitätsspekulationen, was jedoch mit erheblichen Risiken verbunden ist. Die Philosophie kann hier wertvolle Impulse liefern, um neue Trainingsansätze zu

entwickeln und die Leistungsfähigkeit von AI-Modellen im Bereich der Argumentation und Erklärung zu verbessern.

### **8.3.3 Identifikation und Beurteilung von Argumenten als zentrale Aufgabe der AI**

Die Identifikation und Beurteilung von Argumenten ist eine der wichtigsten Aufgaben, die AI-Modelle beherrschen müssen. Argumente sind die Angabe von Gründen, die die Plausibilität oder Wahrheit einer These stützen, erhöhen oder belegen. Sie bilden den Kern jeder Erklärung und sind damit von zentraler Bedeutung für die Beantwortung von Fragen wie “Warum ist etwas so, wie es behauptet wird?” oder “Warum ist etwas nicht so oder wie ist es zu kritisieren?”.

### **8.3.4 Philosophische Befunde zur Argumentation**

Die Philosophie hat sich intensiv mit der Struktur und Beurteilung von Argumenten beschäftigt und kann hier wertvolle Ergebnisse vorweisen. Diese Erkenntnisse müssen nun für die Weiterentwicklung von AI-Modellen nutzbar gemacht werden. Dabei geht es zunächst darum, AI-Modellen beizubringen, Argumente überhaupt zu erkennen und deren Struktur zu analysieren.

### **8.3.5 Unterstützung von AI-Modellen bei der Argumentanalyse**

Mit der richtigen Hilfestellung können AI-Modelle lernen, Argumente zu identifizieren und zu beurteilen. Dazu bedarf es keiner übermäßig komplexen Ansätze, sondern grundlegender Kenntnisse aus der Philosophie und Logik. Durch gezielte Anleitung und Fokussierung auf die wesentlichen Aspekte eines Textes können AI-Modelle Schritt für Schritt an die Analyse von Argumenten herangeführt werden.

#### **8.3.5.1 Formulierung des Erkenntnisinteresses**

Ein wichtiger Schritt dabei ist die Formulierung des Erkenntnisinteresses an einem Text. Im Falle der Argumentation lautet die zentrale Frage: Ist das präsentierte Argument schlüssig? Diese philosophische Kernfrage muss jede AI beherrschen, um sinnvolle Erklärungen und Argumentationen zu liefern.

#### **8.3.5.2 Identifikation von Argumenten und Beurteilung ihrer Schlüssigkeit**

Die zwei Kernelemente, die jede AI in Bezug auf Argumentation beherrschen muss, sind:

1. Die Identifikation von Argumenten in einem Text
2. Die Beurteilung der Schlüssigkeit dieser Argumente

Nur wenn eine AI in der Lage ist, diese beiden Aufgaben zuverlässig zu bewältigen, kann sie als wirklich intelligent bezeichnet werden. Dazu bedarf es einer gezielten Schulung und der Integration philosophischer Erkenntnisse in die Entwicklung von AI-Modellen.## Argumentanalyse in der Philosophie

Meine Damen und Herren, ich möchte heute mit Ihnen über ein faszinierendes Thema sprechen: die Beurteilung von Argumenten mithilfe von AI-Systemen. Stellen Sie sich vor, Sie haben eine Frage und möchten wissen, ob es in Ihrer umfangreichen Bibliothek mit tausenden von Büchern oder in einer Sammlung von Expertenpublikationen ein Argument gibt, das eine bestimmte These begründet. Diese Fragestellung ist der Kern jeder erklärungsbedürftigen Suche.

### **8.3.6 Ein Schema der Argumentanalyse**

Das Schema für eine solche Anfrage ist im Grunde gar nicht so kompliziert. Sie können es in jedem Chat selbst ausprobieren und sehen, was zunächst mit Voreinstellungen herauskommt. Doch die Zukunft liegt darin, diese Voreinstellungen durch verfeinerte Expertenkomponenten zu ersetzen.

Lassen Sie uns dieses Schema genauer betrachten, als ob wir uns in einer mentalen Chatsituation befinden und fehlende Informationen hinzufügen. Wir haben eine Instruktion, die eine auszuführende Handlung formuliert. Obwohl es hier um eine Maschine geht und Handlungen normalerweise personenbezogen sind, erwarten wir, dass das AI-System diese Handlung ausführt und zu einem Ergebnis führt.

#### **8.3.6.1 Detaillierung des Schemas**

Detaillierter betrachtet, besteht unser Schema aus folgenden Elementen:

- Eine Instruktion, die sich auf einen Text bezieht und ein Ziel formuliert (hier: Beurteilung der Schlüssigkeit eines Arguments)
- Die Ausführung durch das KI-System
- Das erwartete Ergebnis (hier: das Urteil über die Schlüssigkeit des Arguments)

Wir erwarten eine korrekte Ausführung dieser Instruktion, ohne halluzinierte oder schlecht begründete Antworten.

### **8.3.7 Die Kernfrage**

Die Kernfrage lautet: Gibt es derzeit ein Verfahren, das bei der Eingabe einer sehr großen Menge von Sätzen (z.B. Millionen) entscheiden kann, ob diese Sätze ein Argument enthalten, das eine zu beurteilende These rechtfertigt? Und zwar ein Verfahren, das auch jetzt ausführbar ist?

Dank der Erträge der Philosophie der letzten 100 Jahre können wir diese Frage prinzipiell bejahen - bei beliebig großen, aber endlichen Mengen von Sätzen. Die Zahl der Sätze ist nicht so entscheidend. Entscheidend sind die philosophischen Komponenten, die man braucht, um eine solche prinzipielle Frage an einem prinzipiellen Text grundsätzlich zu entscheiden.

### **8.3.8 Zuverlässigkeit der Argumentanalyse durch AI**

Was Sie derzeit von AI-Modellen erhalten, sind oft nur Meinungsäußerungen, deren Korrektheit der normale Nutzer nicht beurteilen kann. Darauf sollten Sie Ihre eigenen wissenschaftlichen Arbeiten nicht stützen.

Wenn Sie jedoch sicher sind, dass es ein Verfahren gibt, bei dem die Beantwortung der Frage nach der Bewertung von Argumenten prinzipiell stimmt, haben Sie einen Schlüssel in der Hand. Es ist wie bei einem Taschenrechner: Sie erwarten zuverlässige Ergebnisse, weil Sie wissen, dass er korrekt konstruiert ist.

Diese Zuverlässigkeit erwarten wir jetzt auch von der KI im Bereich der Argumentanalyse. Die Philosophie hat alle nötigen Techniken und Instrumente dafür. Es wird wohl nur noch ein halbes bis ein Jahr dauern, bis die Modelle diese Techniken umsetzen.

## **8.4 Ein Beispiel für Argumentanalyse**

Betrachten wir ein Beispiel, um zu sehen, wie AI-Modelle derzeit mit der Aufgabe der Argumentanalyse umgehen. Dazu geben wir folgenden Text ein:

[Relevantes Argument] Lebensstandard in Indien ist nicht so hoch wie in Italien. [Irrelevanter Satz im Kontext]

Wir instruieren das Modell, den Text danach zu beurteilen, ob ein schlüssiges Argument vorliegt oder nicht, und sich dabei auf die Beurteilung des Arguments selbst zu konzentrieren, nicht auf die Sachbeurteilung.

Das beste aktuelle Modell kommt zu dem falschen Befund, dass kein schlüssiges Argument vorliegt. Das darf nicht passieren - es ist, als würde ein Taschenrechner grundsätzlich falsche Zahlen ausgeben. Die Logik ist sehr eindeutig, was ein Argument ausmacht und was es heißt, ein Argument auf Schlüssigkeit zu prüfen.

#### **8.4.1 Schlüssigkeit vs. Triftigkeit**

Für Philosophen ist der Unterschied zwischen Schlüssigkeit und Triftigkeit klar:

- Ein Argument ist schlüssig, wenn unter der Voraussetzung der Prämissen die Konklusion mit Notwendigkeit folgt, unabhängig von der Wahrheit der Prämissen.
- Triftigkeit hingegen erfordert zusätzlich, dass die Prämissen faktisch wahr sind.

Das Modell hat hier offenbar ein Sprachmissverständnis. Es verwendet den Begriff "Gültigkeit" im Sinne der philosophischen Definition von Schlüssigkeit. Solche Missverständnisse können wir durch Interaktion in Chats korrigieren, indem wir die Begriffe präzisieren.

Nach der Korrektur kommt das Modell zu dem richtigen Befund, dass das präsentierte Argument logisch schlüssig ist. Die Beurteilung ist nun klar.

#### **8.4.2 Methodisches Vorgehen bei der Argumentanalyse**

Um die Frage der Schlüssigkeit eines Arguments noch präziser zu beantworten, können wir dem Modell eine detaillierte Instruktion geben, wie es methodisch vorgehen soll. Diese Methode formulieren wir in üblicher Sprache und geben zusätzlich sprachklärende Definitionen.

Die Modelle arbeiten im Hintergrund mit zwei prinzipiellen Verfahren:

1. Eine lineare Abfolge von Handlungen (Verkettung von Arbeitsschritten)
2. Eine Aufgliederung der Aufgabe in Teilaufgaben

Beide Elemente sind kritisch für die Lösung der ursprünglichen Frage. Welche Aufteilung von Arbeitsschritten wird geplant? Und wie kann eine komplexe Aufgabe in Teilaufgaben zergliedert werden?

Wenn ein Modell gut darin ist, einen solchen Handlungsplan zu entwickeln, dann kann es - wie die Philosophie beweisen kann - die Frage nach der Schlüssigkeit eines Arguments prinzipiell korrekt beantworten. Dann erhalten wir von den KI-Modellen eine Garantie für die Korrektheit der Beurteilung, ähnlich wie bei einem Taschenrechner.

### 8.4.3 Ein Verfahren nach Wittgensteins Tractatus

Ich möchte Ihnen nun ein Verfahren vorstellen, das auf Wittgensteins Tractatus zurückgeht. Dieses Verfahren ist für unsere Zwecke vollkommen ausreichend und hat die gewünschte Qualität: Wenn man es korrekt umsetzt, erhält man eine Garantie für die Korrektheit der Argumentbeurteilung.

Die Details dieses Verfahrens werde ich Ihnen in der nächsten Vorlesung erläutern. Bis dahin können Sie selbst ein wenig experimentieren und sehen, wie weit Sie mit den derzeitigen AI-Modellen bei der Analyse von Argumenten kommen.

Ich freue mich darauf, dieses spannende Thema in der nächsten Sitzung mit Ihnen zu vertiefen. Vielen Dank für Ihre Aufmerksamkeit!## Einleitung in die Methode der Wahrheitstabellen

In der heutigen Vorlesung möchte ich Ihnen eine mächtige Methode vorstellen, mit der wir die Schlüssigkeit von Argumenten überprüfen können - die Methode der Wahrheitstabellen, auf Englisch auch als "Truth Tables" bekannt. Diese Methode geht zurück auf den österreichischen Philosophen Ludwig Wittgenstein und sein bahnbrechendes Werk, den "Tractatus Logico-Philosophicus".

Wittgenstein postulierte, dass man bei einer präzisen Reformulierung von Sätzen garantieren kann, dass die Antwort auf die Frage nach der Gültigkeit eines Arguments immer korrekt ist. Wie funktioniert das nun im Detail? Lassen Sie es mich Ihnen Schritt für Schritt erklären.

### 8.4.4 Anwendung der Methode auf ein konkretes Beispiel

Um die Methode der Wahrheitstabellen anzuwenden, geben wir zunächst ein Argument ein, das wir auf seine Schlüssigkeit hin überprüfen wollen. Dazu verwenden wir eine spezielle Notation: "ARG" steht dabei als Platzhalter für einen beliebigen Text, den wir analysieren möchten. Das kann ein philosophischer Text sein, ein literarischer Text wie Goethes Briefe, oder sogar ein juristischer Text.

Nun fügen wir diesen Text in unser Analyseschema ein und starten einen neuen Chat mit unserem AI-Modell. Wir geben den Begriff "Truth Table" ein und warten gespannt, was das Modell damit anfängt.

## 8.5 Analyse mit Wahrheitstabellen

Das AI-Modell beginnt nun, das Argument mit Hilfe von Wahrheitstabellen zu analysieren, um zu bestimmen, ob es gültig ist oder nicht. Dabei verwendet es den Begriff "gültig" anstelle von "schlüssig", aber die Bedeutung ist dieselbe.

### 8.5.1 Identifikation der Aussagen im Text

Der erste Schritt in diesem Verfahren ist es, die im Text enthaltenen Aussagen zu identifizieren. Dies ist ein entscheidender Schritt, denn wenn hier Fehler passieren, wird alles Folgende problematisch.

In einem beliebigen Text ist nämlich zunächst einmal nicht klar, ob bestimmte Ausdrücke überhaupt Sätze sind, und wenn ja, welche Art von Sätzen. Für unsere Analyse benötigen wir Propositionen, also Sätze, die Aussagen ausdrücken, die entweder wahr oder falsch sind. Nicht jeder Satz erfüllt dieses Kriterium - Ausrufe, Befehle, Wünsche etc. sind keine Propositionen.

In unserem Beispieltext identifiziert das Modell folgende vier Propositionen:

- P: Die Menschheit erzeugt zu viel CO<sub>2</sub>.
- Q: Der Wasserspiegel des Ozeans steigt.
- R: Der Lebensstandard in Italien ist sehr hoch.
- S: Der Lebensstandard in Indien ist so hoch wie in Italien.

### 8.5.2 Berechnung der Wahrheitswerte

Im nächsten Schritt berechnet die Wahrheitstafel eine Kombination aller möglichen Wahrheitswerte für diese vier Propositionen. Bei vier Variablen ergeben sich 16 Zeilen, die alle Kombinationen von "wahr" und "falsch" durchspielen.

Die zu analysierenden Argumente lauten dabei:

1. Wenn P, dann Q.
2. R und P.
3. Nicht S, also Q.

Das Modell erkennt diese Argumente korrekt und trägt sie in die Wahrheitstabelle ein.

### 8.5.3 Fehleranalyse und Korrektur

Nun passiert jedoch etwas Überraschendes: Obwohl wir die "wasserdichte" Methode Wittgensteins anwenden, kommt das Modell zu dem Schluss, dass das Argument nicht gültig ist. Wie kann das sein?

### **8.5.3.1 Fehlersuche in der Wahrheitstabelle**

Um den Fehler zu finden, müssen wir die Wahrheitstabelle genau unter die Lupe nehmen. Ein Argument ist genau dann gültig, wenn in jeder Zeile, in der alle Prämissen wahr sind, auch die Konklusion wahr ist.

Das Modell behauptet nun, es gäbe eine Zeile (nämlich die zweite), in der alle Prämissen und die Konklusion wahr sind, in allen anderen Zeilen mit wahren Prämissen sei die Konklusion falsch. Doch dieser Befund ist nicht korrekt.

Tatsächlich gibt es gar keine Zeile in der Tabelle, in der alle drei Prämissen wahr sind und die Konklusion falsch ist - ein solcher Fall existiert schlichtweg nicht. Hier ist dem Modell offenbar ein simpler Zählfehler unterlaufen.

### **8.5.3.2 Korrektur durch präzise Anweisungen**

Um solche Fehler zu vermeiden, müssen wir dem Modell präzise Anweisungen geben, wie es vorzugehen hat. Wir trainieren es sozusagen philosophisch, indem wir zentrale Begriffe vorgeben und die nötigen Schritte explizit erklären.

Dazu gehören Anweisungen wie:

1. Standardisiere die Ausdrücke, um Missverständnisse zu vermeiden.
2. Führe eine formallogische Analyse durch.
3. Erstelle eine Tabelle mit den Sätzen und ihrer logischen Struktur.
4. Erläutere Begriffe wie "Annahme", "Konklusion" etc.
5. Erkläre, wie die Schlüssigkeit geprüft wird.
6. Streiche irrelevante Zeilen und fokussiere auf die entscheidenden Fälle.

Wenn wir diese Schritte befolgen und an den richtigen Stellen nachjustieren, können wir die begrifflichen Lücken im Modell schließen und erhalten ein Verfahren, das zuverlässig die Schlüssigkeit von Argumenten überprüft.

### **8.5.4 Ausblick auf weiterführende Methoden**

Die hier vorgestellte Methode der Wahrheitstabellen ist ein erster, grundlegender Schritt. Die analytische Philosophie hat seit den 1930er Jahren noch effizientere Verfahren entwickelt, um die Schlüssigkeit von Argumenten zu überprüfen.

Ein Beispiel dafür ist der Quine-McCluskey-Algorithmus, der die Tabellenverfahren abkürzt und optimiert. Auch diesen Algorithmus können wir in unser AI-Modell integrieren, sodass es auf Anfrage den Algorithmus ausgibt und anwendet.

All diese Methoden und Hintergründe sind in den AI-Modellen bereits angelegt - es gilt nur, die richtigen begrifflichen Lücken zu füllen und präzise Anweisungen zu geben. Dann haben wir ein mächtiges Werkzeug, um schlüssig zu argumentieren und zu schließen - die Basis für alle weiteren Kompetenzen, die wir in Zukunft behandeln werden.

## 9 Akteure

Sehr geehrte Damen und Herren,

ich freue mich, Sie heute zur achten Vorlesung unserer Reihe “Philosophie der AI” begrüßen zu dürfen. Heute möchte ich Ihnen das spannende Thema “AI-Akteure” vorstellen. Im Englischen wird der Begriff “Akteur” häufig mit “Agent” übersetzt. Da diese Übersetzung zumindest in Berlin eher Erinnerungen an den Kalten Krieg weckt als an künstliche Intelligenz, werde ich im Folgenden den Begriff “Akteur” verwenden. Ein AI-Akteur kann als eine Art handelnde künstliche Persönlichkeit verstanden werden, auch wenn es sich dabei natürlich nicht um eine menschliche Person, sondern um ein künstliches Intelligenzmodul handelt.

### 9.1 Was ist ein AI-Akteur und wie wird er gebaut?

In unserer heutigen Vorlesung werden wir ergründen, was genau ein AI-Akteur ist, wie er entwickelt wird und welche Kompetenzen er benötigt, um die vielfältigen Funktionen zu erfüllen, die wir gleich näher betrachten werden. Dazu möchte ich Ihnen zunächst unser Projekt “Magister Faustus” vorstellen. In diesem Projekt, dessen Details ich Ende nächster Woche online stellen werde, geht es darum, mit Hilfe von AI-Akteuren die Biografie von Johann Wolfgang von Goethe auf Basis der Goethe-Quellen der Stiftung Weimarer Klassik zu erforschen.

#### 9.1.1 Sprach- und Bedeutungskompetenz als Basis

Wie gehen wir dabei vor? Die Basis bildet natürlich die beeindruckende Sprachkompetenz, die aktuelle AI-Modelle durch ihr Training bereits mitbringen. Sie sind in der Lage, Sprache nicht nur in Form von Symbolen zu verarbeiten, sondern tatsächlich die Bedeutungen der Ausdrücke zu erfassen. Das ist eine enorme Leistung und die Grundlage für die derzeitige Revolution im Bereich der künstlichen Intelligenz.

### **9.1.2 Wissen, Kompetenzen und Charakter**

Doch Sprachkompetenz allein reicht bei weitem nicht aus. Wie wir bereits erforscht haben, fehlt es den aktuellen Modellen noch an Wissen und Kompetenzen in vielen anderen Bereichen sowie an spezifischen Charaktereigenschaften. Diese zusätzlichen Aspekte müssen wir definieren und den AI-Akteuren beibringen.

## **9.2 Apples Umsetzung von AI-Akteuren**

Am Montag hat Apple auf seiner Entwicklerkonferenz seine eigene Umsetzung von AI-Akteuren vorgestellt. Was Sie hier auf meinem Rechner sehen, ist bereits die Developer-Version des neuen Apple-Betriebssystems. Die angekündigten Komponenten zur Sprachverarbeitung und -kompetenz sind darin allerdings noch nicht implementiert.

Dennoch möchte ich kurz skizzieren, warum ich glaube, dass wir uns mitten in einer technologischen Revolution befinden, die alle Bereiche des Lebens - Schule, Hochschule, Politik und Alltag - massiv verändern wird. Lassen Sie mich dazu einige Punkte aus der Apple-Präsentation herausgreifen:

- Ende dieses Sommers sollen die KI-Modelle frei zugänglich sein - für jeden, der ein aktuelles Apple-Gerät mit dem neuen Betriebssystem nutzt, egal ob iPad, MacBook Air, MacBook Pro oder andere Geräte auf Basis des modernen "Apple Silicon" Prozessors. Ab Winter wird dies sogar für iPhones möglich sein.
- Ein interessanter Aspekt ist die Ankündigung, dass ChatGPT 4.0 kostenlos sein wird. Jeder Nutzer eines entsprechenden Geräts wird die Funktionen, die wir hier studieren, kostenlos nutzen können. Ein kluger Schachzug, um die Verbreitung und Nutzung in allen Bereichen zu fördern.
- Aktuell nutzen bereits 2,2 Milliarden Menschen weltweit Apples Geräte - eine enorme potenzielle Nutzerbasis für die neuen AI-Funktionen.

### **9.2.1 Multimediale Verarbeitung und tiefe Integration**

Die Version, die man heute schon nutzen kann und die ich gleich für Demonstrationen verwenden werde, ist multimedial. Sie verarbeitet Text, Bilder, Audio und die gängigsten Videoformate mit Hilfe von künstlicher Intelligenz. All diese KI-Einheiten sind tief in die neuen Betriebssysteme integriert.

Wenn Sie beispielsweise E-Mails, PDFs oder Kalendereinträge auf Ihrem Gerät für den Zugriff durch die KI freigeben, kann der AI-Akteur direkt darauf zugreifen und die Informationen auswerten. Stellen Sie sich vor, Sie erhalten eine E-Mail, in der erwähnt wird, dass eine bestimmte Person in einem Monat Geburtstag hat. Sie könnten Ihr iPhone dann per Sprachbefehl anweisen: "Suche ein passendes Konzert als Geschenk für diese Person an ihrem Geburtstag." Basierend auf den freigegebenen Informationen zu Ihrem Verhältnis zu dieser Person, ihren musikalischen Vorlieben usw. würde der AI-Akteur die Aufgabe analysieren und ein perfekt passendes Geschenk vorschlagen. Beeindruckend, finden Sie nicht?

### **9.2.2 Datenschutz und Privatsphäre**

Natürlich geht damit auch die Sorge einher, ob und inwieweit die Privatsphäre dieser hochgradig persönlichen Informationen, die man dafür freigeben muss, geschützt ist. Apple hat in der Präsentation und in verschiedenen Workshops ausführlich dafür geworben, dass zwei technische Aspekte die Privatsphäre der Nutzerdaten sicherstellen:

1. Die weniger rechenintensiven KI-Anwendungen laufen direkt auf dem Gerät selbst, wofür die schnellen und leistungsfähigen "Apple Silicon" Prozessoren benötigt werden. Der Großteil der Informationsverarbeitung findet also lokal auf dem eigenen Gerät statt.
2. Für anspruchsvollere Berechnungen werden die Daten in einer speziellen, vollständig verschlüsselten Cloud verarbeitet. Diese Verschlüsselung soll so sicher sein, dass weder Apple selbst noch Sicherheitsbehörden darauf zugreifen können.

Ohne einen solchen Schutz der Privatsphäre wäre die Akzeptanz und Nutzung der neuen Möglichkeiten sicher deutlich eingeschränkt.

### **9.3 Philosophische Konzepte hinter AI-Akteuren**

Lassen Sie uns nun zu den Funktionen selbst kommen. Die KI-Modelle basieren auf einem Konzept, das in der Philosophie als "Handlungsmodell" bezeichnet wird. Dabei geht es um den Zusammenhang zwischen zwei wesentlichen Komponenten: dem Ziel einer Handlung und dem "Belief", also den Überzeugungen und dem Hintergrundwissen, auf deren Basis die Handlung ausgeführt wird, um dem Ziel näher zu kommen oder es zu erreichen.

### 9.3.1 Informationsquellen und Handlungsmodell

Zur Umsetzung dieses Handlungsmodells werden verschiedene Informationsquellen wie Texte, Bilder, Audio, Video und Nutzerdaten ausgewertet, die das Hintergrundwissen für den AI-Akteur bilden. Lassen Sie mich das anhand einer konkreten Anwendung demonstrieren.

Ich wechsele jetzt zu einer App, die im Wesentlichen ChatGPT ähnelt, aber nicht über die Webseite, sondern als eigenständige Anwendung läuft und sukzessive erweitert wird. Die von mir beschriebenen Konfigurationsmöglichkeiten und Erweiterungen sind hier noch nicht implementiert, aber das zugrundeliegende Konzept ist dasselbe.

Im Gegensatz zu unserem Vorgehen in der ersten Vorlesung müssen wir jetzt keine Texte mehr eingeben, hochladen oder tippen. Stattdessen können wir, wie im täglichen Umgang mit Computern üblich, einfach Informationen in einem Bild bereitstellen. Hier sehen Sie ein Beispiel dafür.

Die Aufgabenstellung habe ich der Einfachheit halber in Englisch formuliert, sie funktioniert aber genauso auf Deutsch. Es geht darum, die Instruktionen und das anzuwendende Verfahren so präzise und unmissverständlich zu formulieren, dass der AI-Akteur die Aufgabe exakt wie gewünscht ausführt.

Die meisten Modelle sind primär anhand englischer Texte trainiert, es gibt aber auch Modelle, die zusätzlich auf die acht wichtigsten europäischen Sprachen und insgesamt über 120 Sprachen und Varianten wie Katalanisch spezialisiert sind.

In meinem Beispiel habe ich die Anweisungen auf Englisch formuliert, was aber der einzige Grund für die Verwendung von Englisch ist. Lassen Sie mich die Struktur der Instruktionen für das Handlungsmodell auf Deutsch zusammenfassen:

- Zuerst wird das zu erreichende Ziel möglichst klar und präzise formuliert.
- Dann wird beschrieben, wie dieses Ziel erreicht werden soll, also die anzuwendende Methode.
- Schließlich werden die erforderlichen Hintergrundinformationen spezifiziert, die als Wissensgrundlage dienen.

Der Charme dieser KI-Modelle besteht darin, dass all dies in natürlicher Sprache geschieht - allerdings nicht umgangssprachlich, sondern möglichst präzise formuliert. Hier hat die Philosophie mit ihrer spezifischen Sprache und Methodik gewisse Vorteile.

Anders als in der letzten Woche erhält das Modell hier einen zusammenhängenden Textblock als Input. Ich gebe nicht an, wie dieser Text strukturiert ist. In der Instruktion sage ich lediglich, dass es sich um einen Text handelt, der aus mehreren Sätzen besteht.

Die eigentliche Aufgabe für das Modell besteht darin, zu bestimmen, ob dieser Text ein schlüssiges Argument enthält oder nicht. Dazu muss es zunächst die These identifizieren, für die in dem Text argumentiert wird. Anschließend muss es prüfen, ob die weiteren Sätze geeignet sind, die Wahrheit oder Falschheit dieser These zu begründen. Das ist eine anspruchsvolle inhaltliche Aufgabe.

Wie wir letzte Woche gesehen haben, liefern die standardmäßig trainierten Modelle einschließlich ChatGPT bei ungenauen Instruktionen oft keine zufriedenstellenden Ergebnisse. Ihnen fehlt noch die Anreicherung durch streng methodische, philosophische Verfahren. Deshalb habe ich das anzuwendende Verfahren hier in vier präzisen Schritten formuliert, die ich gegenüber der letzten Vorlesung noch weiter verfeinert habe.

Hier zeigt sich eine Besonderheit von AI-Akteuren: Genau wie bei menschlichen Akteuren ist eine Aufgabe meist nicht mit einem einzigen Schritt zu lösen, sondern erfordert eine Abfolge verschiedener Teilschritte. Denken Sie nur an die Lösung eines wissenschaftlichen Problems, für die Forscher oft ein ganzes Jahr im Labor arbeiten - von den Experimenten über die Auswertung bis hin zur Publikation der Ergebnisse in einem Journal. All diese Arbeitsschritte würden von einem AI-Akteur ausgeführt. Selbst wenn es sich dabei um Tausende einzelner Schritte handelt, stellt das für die KI kein Problem dar. Komplexität und Umfang sind für sie die geringsten Hürden.

Die von uns untersuchte Prüfung der Validität von Argumenten, also deren logische Schlüssigkeit, bildet den Kern jeder Bewertung von Hypothesen und Theorien. Sie ist essenziell für die Analyse der empirischen Basis, für die Prüfung von Kritik und Gegenargumenten und für die Falsifikation von Hypothesen. Die Grundlage dafür bilden die logischen Schlussverfahren, von denen wir hier die einfachste Variante, die Aussagenlogik, betrachten.

Interessanterweise lässt sich dieses Verfahren dem KI-Modell anhand von vier einfachen, umgangssprachlich formulierten Anweisungen beibringen. Das geht weit über die reine Sprachkompetenz hinaus. Lassen Sie mich die einzelnen Schritte kurz vorstellen, ohne jetzt auf jedes Detail einzugehen. Sie werden sehen, dass dahinter keine Magie, sondern durchsichtige Logik und elementare philosophische Methodik steckt.

1. Unterteile den Text in nummerierte Sätze.

- Hier geht es darum, dem Modell zu vermitteln, was die Voraussetzungen für ein Argument sind: die Prämissen, die Konklusion, die These selbst. Das muss explizit gemacht werden. In der Philosophie ist der natürliche Ausgangspunkt dafür, mit Sätzen zu arbeiten, die wahr oder falsch sein können und Propositionen ausdrücken. Der erste Schritt besteht also darin, den Text in solche Sätze zu zerlegen.

Im Einzelfall kann das durchaus komplex sein, etwa wenn ganze Bücher oder Bibliotheken analysiert werden sollen. Stellen Sie sich vor, die Aufgabe lautet: Gibt es in der gesamten Bibliothek der Humboldt-Universität auch nur ein einziges Buch, das meiner These P widerspricht? Das ist manuell nicht zu bewältigen, für die KI aber durchaus lösbar. Um eine solche Aufgabe zu meistern, kann man das hier skizzierte Verfahren anwenden.

Beachten Sie, wie elegant und schlank dieses Verfahren ist. Mehr als die hier aufgeführten Schritte sind nicht erforderlich. Natürlich muss bei einem so umfangreichen Beispiel mit ganzen Bibliotheken noch etwas Infrastruktur bereitgestellt werden...## Einleitung in die Verarbeitung natürlicher Sprache mit AI-Modellen

In dieser Vorlesung möchte ich Ihnen zeigen, wie man mit modernen AI-Sprachmodellen natürliche Texte analysieren und logische Argumente auf ihre Schlüssigkeit überprüfen kann. Die Skalierung solcher Modelle stellt heutzutage kein Problem mehr dar. Lassen Sie mich Ihnen nun Schritt für Schritt erklären, wie man dabei vorgeht.

### **9.3.2 Zerlegung des Textes in Einzelsätze und Standardisierung**

Zunächst werden alle zu analysierenden Texte segmentiert und in einzelne Sätze zerlegt. Im nächsten Schritt werden diese Sätze standardisiert, sodass die Verbindungen mit logischen Operatoren wie “und”, “oder”, “nicht”, “wenn...dann” einzelne Teilsätze miteinander verknüpfen. Hierbei bilden wir sogenannte propositionale Funktionen. Diese Begriffe sollten unbedingt verwendet werden, da die Modelle speziell dafür trainiert sind.

Es empfiehlt sich, verschiedene Formulierungen auszuprobieren - auch in Ihrer eigenen Sprache - um herauszufinden, ob das Programm die Instruktionen und das beschriebene Verfahren versteht. Falls nicht, müssen Sie Ihre Formulierungen so anpassen, dass das Sprachtraining die entsprechenden Begriffe erkennt.

### **9.3.3 Zuweisung von Buchstaben zu Aussagen gleicher Bedeutung**

Ein wichtiger Schritt, den ich durch Ausprobieren herausgefunden habe und der nicht von vornherein vom Modell ausgeführt wurde, ist die Zuweisung des gleichen Buchstabens zu Aussagen mit derselben Bedeutung. Dies ist eine wunderbare Instruktion, die vor zwei Jahren noch nicht so einfach möglich war - egal wie leistungsfähig die Computer waren. Heute reicht es aus, einen Text als Methode zu formulieren, ohne auf technische Tricks zurückgreifen zu müssen.

### **9.3.4 Umsetzung des Arguments in eine kompakte Form**

Im nächsten Schritt geht es darum, das Argument in eine kompakte Form zu überführen, an der man direkt ablesen kann, ob es schlüssig ist oder nicht. Dazu bilden wir einen neuen Ausdruck, in dem alle Voraussetzungen oder Prämissen durch eine "und"-Verknüpfung verbunden werden. Die Schlussfolgerung (Konklusion) aus diesen Prämissen wird dann über eine "wenn...dann"-Konstruktion dargestellt: Wenn die Voraussetzungen erfüllt sind, dann gilt die Konklusion. Ein einfaches Standardverfahren aus der Logik, das ich hier nochmals explizit als Regel formuliert habe.

### **9.3.5 Anwendung von Wahrheitstabellen zur Überprüfung der Schlüssigkeit**

Der letzte Kniff, den ich in der vergangenen Stunde vorgestellt habe, ist ein alter Trick von Wittgenstein: ein Verfahren, das in diesem Fall mechanisch immer zu einer Lösung führt. Das Schöne daran ist, dass es nicht mit Wahrscheinlichkeiten arbeitet und nicht diskutiert werden kann. Das Verfahren ist eindeutig und liefert ein klares Ergebnis.

Bei endlichen Aussagen wenden wir Wahrheitstabellen an, in denen alle mit Buchstaben gekennzeichneten einfachen Aussagen mit Wahr- und Falsch-Werten kombiniert werden. Anschließend wird berechnet, ob der Gesamtausdruck des Arguments unabhängig von den Wahrheitswerten der einzelnen Aussagen wahr ist oder nicht. Ist dies der Fall, so ist das Argument schlüssig - ganz eindeutig und unmissverständlich. Das Programm muss dies nur entsprechend ausführen.

## **9.4 Demonstration des Verfahrens mit ChatGPT 4.0**

Nun möchte ich Ihnen das beschriebene Verfahren anhand von ChatGPT 4.0 demonstrieren - jener Version, die Apple im Herbst kostenlos allen Nutzern zur Verfügung stellen wird. Jeder Besitzer eines Apple-Geräts wird darauf zugreifen können, wobei ich davon ausgehe, dass auch andere Anbieter wie Microsoft mit ihrem KI-Projekt OpenAI ähnliche Implementierungen für Windows-Systeme anbieten werden. Ob und inwieweit dies auch für Linux gelten wird, bleibt abzuwarten.

In der heutigen Zeit geht es nicht mehr darum, komplexe Programme zu schreiben. Stattdessen kommt es darauf an, präzise zu formulieren - eine spannende Herausforderung gerade für Philosophen.

### 9.4.1 Beispieltext und Fragestellung

Betrachten wir nun folgenden Beispieltext, den wir bereits in einer früheren Sitzung diskutiert haben und der einige von Ihnen möglicherweise etwas verwirrt hat:

- Wenn die Menschheit zu viel CO<sub>2</sub> produziert, wird der Wasserspiegel der Ozeane ansteigen.
- Der Lebensstandard in Italien ist sehr hoch und die Menschheit produziert zu viel CO<sub>2</sub>.
- Der Lebensstandard in Indien ist nicht so hoch wie der in Italien.
- Heute finden Wahlen in Indien statt.

Die Schlussfolgerung lautet: “Deshalb wird der Wasserspiegel der Ozeane steigen.”

Die Frage, die sich uns stellt, ist: Handelt es sich hierbei um ein schlüssiges Argument? Manche lassen sich von den irrelevanten Zwischenannahmen irritieren und argumentieren, dass diese nichts mit der Konklusion zu tun haben. Das stimmt zwar, aber das Hinzufügen irrelevanter Informationen zu den Prämissen ändert nichts an der Schlüssigkeit der Schlussfolgerung.

Man könnte das Verfahren noch erweitern und zunächst alle irrelevanten Informationen aus dem zu untersuchenden Text entfernen. Der verbleibende Rest enthielte dann nur noch die für das Argument relevanten Aussagen. Für unsere Zwecke werden wir dies jedoch nicht tun, sondern prüfen, ob das Verfahren in seiner jetzigen Form wasserdicht ist.

### 9.4.2 Anforderungen an AI-Systeme

Von KI-Systemen und Machine-Learning-Implementierungen wird heutzutage nicht nur erwartet, dass sie ein Ergebnis liefern, sondern auch, dass sie nachvollziehbar erklären können, wie sie zu diesem Ergebnis gelangt sind. Jeder sollte in der Lage sein zu überprüfen, ob das Resultat korrekt ist. Die Akzeptanz des Ergebnisses darf nicht von der zugrundeliegenden Technik, dem Hersteller oder irgendeiner Autorität abhängen - insbesondere dann nicht, wenn die Befunde mit hoher Verantwortung weiterverwendet werden.

## 9.5 Interaktion mit dem AI-Agenten

Lassen Sie mich Ihnen nun zeigen, wie man mit dem AI-Agenten interagiert und welche Informationen diese App verarbeiten kann. Anstatt den Text wie in der letzten Stunde einfach zu kopieren, werde ich diesmal einen anderen Ansatz wählen.

### **9.5.1 Nutzung von ChatGPT und Antwortzeiten**

Zunächst möchte ich anmerken, dass die Nutzung von ChatGPT derzeit in die Millionen geht. Obwohl massive Rechnerkapazitäten dahinterstehen, sind die Antwortzeiten aufgrund der enormen Nachfrage etwas langsamer geworden. Die derzeitige Infrastruktur schafft es nur knapp, die parallel eingehenden Anfragen zeitnah zu bearbeiten. Daher werde ich die Anfrage gleich starten.

### **9.5.2 Der AI-Agent als Diskussionspartner**

Bevor ich dazu komme, möchte ich Ihnen noch verdeutlichen, wie der AI-Agent immer stärker zu einem echten Diskussions- und Austauschpartner wird. Wenn Sie das Kopfhörer-Symbol rechts unten anklicken, wird der Agent aktiv und meldet sich mit einer Frage wie “Kannst du mir mehr Details zu der Szene geben? Möchtest du, dass ich eine Geschichte daraus entwickle oder soll ich etwas Bestimmtes analysieren?”

Meine Antwort darauf könnte lauten: “Als Erstes rede nicht so viel, sondern mache das, was ich sage. Ein Agent wird wach und startet seine Mission.” Daraufhin antwortet der Agent sehr sachlich.

### **9.5.3 Verarbeitung von Bildern und Argumenten**

Als Nächstes frage ich den Agenten: “Kannst du gleich ein Bild verarbeiten, mit dem ein Argument analysiert werden soll und dieses Ergebnis zurückgeben?” Der Agent bestätigt, dass er dazu in der Lage ist und bittet mich, das zu analysierende Bild hochzuladen.

Dieser Dialog mag in diesem Fall noch etwas unergiebig erscheinen, aber er verdeutlicht, wie man mit dem Agenten interagieren und interessante Gespräche führen kann - beispielsweise zur Lösung mathematischer Aufgaben.

Wie bereits erwähnt, sind diese Dialoge ein geschickter Kniff, um über die Interaktion den Nutzer mit seinem Weltwissen und seinen Begrifflichkeiten in die Antworten einzubeziehen. Die Leistung und das Ergebnis einer KI hängen somit nicht nur vom Modell selbst ab, sondern auch von der Interaktion. Daher kann man auch unvollständige Fragen stellen, wie ich es getan habe, als ich den Agenten fragte, ob er überhaupt Bilder verarbeiten kann.

### **9.5.4 Bilderkennung und Wissensgenerierung**

Das Bild, das wir analysieren möchten, ist jenes, das ich in diesem Browserfenster angezeigt habe. Anstelle eines Arguments könnten Sie hier beispielsweise auch ein Foto des Eiffelturms einfügen und den

Agenten fragen, um welches Gebäude es sich handelt und wann es erbaut wurde. Der Agent würde die Fragen in diesem Fall korrekt beantworten, wie ich bereits ausprobiert habe. Für unsere Zwecke bleiben wir jedoch bei unserem ursprünglichen Argument.

Um dem Agenten ein Bild zu übergeben, nutze ich die Funktion "Take a Screenshot". Ich weise den Agenten an, dieses Fenster aufzunehmen, was gerade passiert ist. Das aufgenommene Bild sehen Sie nun in der unteren linken Ecke.

### **9.5.5 Extraktion und Verarbeitung von Textinformationen aus Bildern**

Anstatt den Text aus dem Bild manuell zu extrahieren und in das Textfeld einzugeben, werde ich dem Agenten nun direkt die Anweisung erteilen: "Transkribiere den Text im Bild, interpretiere ihn und führe die darin enthaltene Instruktion aus."

Wie Sie sehen, ist dies ein komplexer Prozess, der weit über das einfache Eingeben eines Befehls hinausgeht. Das Bild muss geladen, der Text extrahiert und korrigiert werden. Anschließend muss die Instruktion verstanden und ausgeführt werden. All dies erfordert eine enorme Informationsverarbeitungsleistung.

Nun übergebe ich die Aufgabe an den Agenten, indem ich auf den Pfeil klicke. Die Transkription des Bildtextes wird ausgegeben und der Agent wiederholt die verstandenen Inhalte, um die Instruktion zu präzisieren. Anschließend wird eine detaillierte Vorgehensweise ausgegeben.

### **9.5.6 Erläuterung der einzelnen Schritte**

Wie von mir gefordert, protokolliert der Agent jeden einzelnen Schritt und gibt eine Erläuterung dazu aus. Leider kann ich die Ausgabe hier nicht vergrößern, aber ich werde versuchen, sie so gut wie möglich sichtbar zu machen.

Der erste Schritt bestand darin, den gesamten Text in einzelne Sätze zu zerlegen. Insgesamt wurden fünf Sätze erkannt und nummeriert. Im zweiten Schritt erfolgte die Standardisierung mit Buchstaben für die Teilsätze. Auch die logische Form wurde korrekt wiedergegeben.

In Schritt C wurden Aussagen mit gleicher Bedeutung demselben Buchstaben zugeordnet. Der Agent bestätigt, dass dieser Schritt erledigt ist und die Bedingung erfüllt wurde. Anschließend wurde in Schritt D eine propositionale Funktion des Arguments erzeugt.

### 9.5.7 Erstellung der Wahrheitstabelle

Nun folgt die Erstellung der Wahrheitstabelle, die ich zuvor erwähnt hatte. Hierbei handelt es sich um eine große Tabelle mit Wahr- und Falsch-Werten. Der eigentliche Rechenaufwand ist schnell erledigt, aber die Übertragung der Tabellen nimmt eine gewisse Zeit in Anspruch.

Abschließend wird geprüft, ob die jeweiligen Sätze, die in den Spalten die komplexen logischen Formen enthalten, so zusammenstehen, dass die Bedingung für schlüssige Argumente erfüllt ist oder nicht. Darauf warten wir jetzt gespannt.## Einleitung in die Nutzung von AI für die Prüfung logischer Argumente

Gestern habe ich einen ersten Versuch unternommen, ein Verfahren zur Prüfung logischer Argumente mit Hilfe eines AI-Akteurs zu entwickeln. Dabei habe ich zwar das Verfahren selbst angegeben, aber versäumt zu erklären, wozu es eigentlich dient. Die AI hat daraufhin pflichtgemäß eine Tabelle erstellt, dann aber ihre Arbeit beendet, ohne den eigentlichen Zweck - nämlich die Bewertung des Arguments als schlüssig oder nicht schlüssig - zu erfüllen.

Heute habe ich diese fehlende Angabe ergänzt und erwarte nun, dass die AI am Ende zu dem Befund kommt, dass das Argument tatsächlich schlüssig ist. Lassen Sie uns gemeinsam das Ergebnis betrachten.

## 9.6 Das Akteurmodell als Grundlage der Interaktion mit AI

Die AI führt nun für jede mögliche Kombination von Wahrheitswerten der Teilaussagen eine Beurteilung durch, ob der gegebene Text ein schlüssiges Argument enthält. Dieser Prozess basiert auf einem Akteurmodell, das vier wesentliche Informationen benötigt:

1. Eine Zielvorgabe: Was soll erreicht oder erstellt werden?
2. Die Informationen, auf deren Basis das Ziel verfolgt wird.
3. Ein möglichst präzises Verfahren, das die einzelnen Schritte zur Zielerreichung beschreibt.
4. Die Angabe, welches Ergebnis kommuniziert werden soll.

All diese Informationen können in Umgangssprache formuliert werden, ohne dass Programmierkenntnisse erforderlich sind. Das zugrundeliegende Schema einer solchen Instruktion ähnelt dem eines Rezepts, wie wir es aus Kochbüchern kennen: Es wird eine Handlungsanweisung gegeben, wie etwas zu tun ist, und am Ende steht bei korrekter Ausführung ein bestimmtes Ergebnis.

### **9.6.1 Die Bedeutung von Instruktionstexten in der Wissenschaft**

Interessanterweise bestehen wissenschaftliche Texte zu einem großen Teil aus solchen rezeptartigen Handlungsanweisungen und nicht, wie man vielleicht vermuten würde, aus Aussagen über wahre und falsche Sachverhalte in der Welt. Man denke nur an pharmazeutische Texte, die beschreiben, wie bei einer bestimmten Krankheit welche Medikamente zubereitet, dosiert und verabreicht werden sollen.

Selbst in klassischen Werken wie Euklids “Elementen” besteht jeder geometrische Beweis aus zwei Teilen: Zunächst wird in einer Art Rezept erklärt, wie ein bestimmtes geometrisches Objekt, etwa ein Dreieck, zu konstruieren ist. Erst im Anschluss daran werden die Eigenschaften dieses Objekts bewiesen, was wir heute als den eigentlichen wissenschaftlichen Beweis ansehen würden.

Auch in modernen experimentellen Studien, seien es Laborexperimente oder Klimasimulationen, macht die Beschreibung des Versuchsaufbaus und der Durchführung einen Großteil des Textes aus. Erst auf dieser Basis werden dann die eigentlichen Ergebnisse präsentiert und diskutiert.

Insofern ist die Sprache der Instruktion keineswegs etwas Neues, sondern ein Kernbestandteil jeder Wissenschaft und auch des Alltagswissens. Das Schema ist immer das gleiche:

- Eine Handlungsinstruktion wird gegeben.
- Diese Instruktion wird ausgeführt.
- Ein Ergebnis wird erzielt und beurteilt, selbst wenn es sich um einen Fehlschlag handelt.

## **9.7 Die Anwendung des Instruktionsschemas auf AI-Akteure**

Dieses Schema lässt sich nahtlos auf AI-Akteure übertragen, die im Prinzip das Gleiche tun wie menschliche Wissenschaftler. Da die Formulierung der Instruktionen weitgehend in Umgangssprache erfolgt, ist der Schritt hin zu einem von AI gesteuerten Wissenschafts- und Laborbetrieb gar nicht mehr so weit.

Wir haben gesehen, dass dieses Prinzip nicht nur für naturwissenschaftliche, sondern auch für geisteswissenschaftliche Forschungsaktivitäten gilt. Wir selbst können solche Instruktionen formulieren, da wir mit der Struktur und Handhabung vertraut sind.

### **9.7.1 Variationsmöglichkeiten der Instruktionen**

Dabei lassen sich die Instruktionen durchaus variieren, etwa indem zusätzliche, für die Schlüssigkeit des Arguments irrelevante Informationen eingefügt werden. Die AI sollte dennoch in der Lage sein, die

wesentlichen Schritte zu extrahieren und korrekt auszuführen.

### **9.7.2 Entwicklungspotenzial der Nutzungsschnittstellen**

Momentan erfolgt die Interaktion mit den AI-Akteuren noch primär über Chat-Schnittstellen, doch ist absehbar, dass sich die Nutzungsweisen und Interfaces rasant weiterentwickeln werden. Neben Chat-GPT gibt es bereits Konkurrenzprodukte wie Claude von Anthropic oder mein eigenes Startup Lettre AI, die zusätzliche Funktionen bieten. Bis zum zweiten Teil dieser Vorlesung im Wintersemester werden wir sicher schon ganz andere Schnittstellen sehen.

## **9.8 Ergebnisse und Perspektiven**

Schauen wir uns nun an, ob die AI ihre Aufgabe zufriedenstellend gelöst hat. Die Tabelle wurde vollständig erstellt, und das System arbeitet noch immer daran, eine abschließende Beurteilung abzugeben. Tatsächlich hat es nun den Text korrekt als schlüssiges Argument eingestuft, ohne dass die Voraussetzungen explizit vorgegeben werden mussten.

Bemerkenswert ist nicht nur die Korrektheit des Ergebnisses, sondern auch die ausführliche Dokumentation der einzelnen Schritte, die praktisch publikationsreif ist. Die AI liefert eine vollständige, nachvollziehbare Erklärung, warum das Fazit gerechtfertigt ist.

### **9.8.1 Perspektiven für den Einsatz von AI**

Dieses Vorgehen eröffnet weitreichende Perspektiven für den zukünftigen Einsatz von AI. Es genügt nicht, sich auf die Ergebnisse einfach zu verlassen - sie müssen überprüfbar sein, sei es durch unabhängige maschinelle Verfahren oder durch manuelle Kontrolle.

Man könnte etwa einen Meta-Agenten entwickeln, der die Korrektheit der angewandten Methodik und die Schlüssigkeit der Ergebnisse überprüft. Dieses Vorgehen wird in vielen Bereichen relevant werden, von der Erstellung von Steuerberichten über rechtliche Beurteilungen bis hin zu technischen Reparaturanleitungen.

Es ist absehbar, dass der Einsatz von AI das Arbeitsleben in Zukunft massiv verändern wird. Die Interaktionsmöglichkeiten werden sich rasant erweitern, etwa durch die Integration von Spracheingabe, das Hochladen von Dateien und Fotos oder die direkte Einbindung von Kamera-Inputs.

### 9.8.2 Ausblick auf die Entwicklung von Instruktionsbibliotheken

Ich erwarte, dass sich schon bald große öffentliche Bibliotheken mit Instruktionen für alle möglichen Verfahren entwickeln werden, sei es für etablierte Standardmethoden oder für individuelle Problemstellungen. Diese Instruktionen werden den Modellen nicht durch Sprachtrainingsdaten beigebracht, sondern als zusätzliche Komponenten oder “Engines” bereitgestellt.

Diese Entwicklung wird weitgehend unabhängig von den jeweiligen Chat-Firmen und ihren Sprachmodellen erfolgen. Die Ausführung der Instruktionen kann je nach Schwierigkeitsgrad auf unterschiedlich anspruchsvolle Modelle verteilt werden, von der lokalen Verarbeitung auf dem Endgerät bis hin zu Cloud-Servern für komplexere Aufgaben.

### 9.8.3 Perspektiven für kollaborative Forschung

Besonders spannend ist die Möglichkeit, die Ausführung der Instruktionen einer großen Community zu überlassen und so viele Akteure an einer gemeinsamen Problemstellung arbeiten zu lassen. Solche kollaborativen Ansätze hat es in der Wissenschaft zwar immer schon gegeben, sie wurden aber eher als Ausnahme wahrgenommen.

Aktuelle Großprojekte, etwa in der Astronomie, zeigen jedoch, dass tausende Akteure gemeinsam an Teilprojekten arbeiten können. Dieses Potenzial lässt sich durch den Einsatz von AI noch erheblich ausweiten und auf viele weitere Forschungsfelder übertragen.## Arbeitsteilige Prozesse in der Wissenschaft

In der heutigen Wissenschaftslandschaft sind arbeitsteilige Prozesse gang und gäbe. Man denke nur an das Paradebeispiel des Hochenergiebeschleunigers am CERN, wo jede einzelne Publikation derzeit noch mit Tausenden von Einzelwissenschaftlern als Autoren benannt wird, die alle ihren Beitrag zu einer Problemstellung leisten. Solche Vorgehensweisen sind sowohl in der Physik als auch in der Medizin Standard und üblich. In Zukunft könnten diese Aufgaben potenziell auch von AI-Akteuren übernommen werden, wobei nichts dagegen spricht, dass nur die Maschine in einer Universität all diese Arbeit leisten sollte.

## 9.9 Auswertung und Validitätsprüfung

Nach der Ausführung einer Handlung folgt die Auswertung. Hierbei geht es nicht nur darum, ein Protokoll der ausgeführten Handlungsanleitung zu erstellen, sondern auch die Validität und Korrektheit des Ergebnisses zu prüfen. Dieser Schritt wird in den jetzigen AI-Modellen oft vernachlässigt oder gar nicht ausgeführt, was ein großer Fehler ist. Wenn wir manuell arbeiten, führen wir diese Prüfung auf die eine

oder andere Weise durch. Nehmen wir das Beispiel einer Rechenaufgabe: Wenn wir eine Zahl als Ergebnis erhalten, lernen wir, dieses Ergebnis unabhängig vom Rechenverfahren auf Korrektheit zu prüfen - ein Standardverfahren, das schon die Babylonier vor 3000 Jahren anwendeten. Wir glauben nicht einfach einem Ergebnis, sondern überprüfen dessen Korrektheit.

### **9.9.1 Ablauf der Planausführung**

Basierend auf dem Prüfungsergebnis wird oftmals ein Auswertungsschritt erforderlich, der festlegt, was als nächstes zu tun ist. Diese Abfolge bildet den Kern der Planausführung. Pläne sind hier als Sequenzen oder Abfolgen von Handlungen zu verstehen, die dazu führen, dass eine ursprünglich große und anspruchsvolle Aufgabe in Teilaufgaben zerlegt wird. Diese Teilaufgaben werden dann nacheinander ausgeführt, ihre Ergebnisse ausgewertet und geprüft. Je nach Prüfungsbefund verzweigt das Vorgehen auf den nächsten Schritt, beispielsweise die Bearbeitung des nächsten Teilschritts.

## **9.10 Standardisierung der Instruktion**

Bevor eine Instruktion ausgeführt werden kann, muss sie zunächst standardisiert werden. Dies erfolgt auf der sprachphilosophischen Ebene, indem man versucht, umfangreiche Textinformationsquellen einschließlich Diktaten in Sätze zu übertragen. Die Bedeutung einzelner Ausdrücke hängt dabei vom Kontext des Satzes ab (Freges Kontextprinzip) oder präziser vom jeweiligen Sprachspiel, also dem Sprachkontext, in dem die Sprachverwendung erfolgt (Wittgensteins Spätphilosophie). Dieser Ansatz wird in den aktuellen Large Language Models implementiert und entsprechend analysiert.

### **9.10.1 Formanalyse und Inferenz**

Nach der Standardisierung folgt die Formanalyse, bei der versucht wird, eine allgemeine Form der Problemstellung zu identifizieren. Dies kann dazu führen, dass komplizierte Texte in Aussagenmengen übertragen werden, auf die dann die Aussagenlogik angewendet werden kann. Darüber hinaus gibt es jedoch noch eine Vielzahl weiterer Module oder Engines, die über die reine Sprachkompetenz hinaus erforderlich sind, um spezifische Anforderungen zu erfüllen. Neben der Form spielt auch die Inferenz eine wichtige Rolle: Welche Schlüsse können aus dem Gegebenen gezogen werden? Hierfür hat die Philosophie seit 2000 Jahren verschiedene Instrumente entwickelt, die jetzt auf die Artificial Intelligence übertragen werden müssen.

### 9.10.2 Epistemische Inferenz

Besonders interessant für die Leistungsfähigkeit von AI-Modellen ist die epistemische Inferenz. “Epistemisch” leitet sich vom griechischen “episteme” ab, was Wissen im Sinne von mathematischem Wissen bezeichnet, in der modernen Sprache aber allgemein für Wissen und nicht Glauben steht. Die epistemische Inferenz umfasst Elemente der Rechtfertigung, Begründung, des Wissensanspruchs sowie der Beziehung zwischen Evidenz und daraus abgeleiteten Konsequenzen - Themen, mit denen sich die Philosophie schon lange beschäftigt. In der KI sind diese Aspekte derzeit noch nicht implementiert, aus meiner Perspektive jedoch von größter Bedeutung.

## 9.11 Logische Verfahren

Zur Formanalyse gehören verschiedene logische Verfahren wie die Satzlogik, die Prädikatenlogik, die Modallogik, das kausale Schließen und die temporale Logik.

- Die Satzlogik haben wir bereits kennengelernt.
- Die Prädikatenlogik zergliedert Sätze in Teilstrukturen und betrachtet deren Innenstruktur, bestehend aus Verben oder Prädikaten, die sich auf benannte Gegenstände beziehen und deren Eigenschaften beschreiben.
- Die Modallogik behandelt Operatoren wie Möglichkeit, Unmöglichkeit oder Notwendigkeit und wird beispielsweise bei der Formulierung von Normen verwendet.
- Das kausale Schließen befasst sich damit, wie Ursachen und Wirkungen miteinander in Bezug gesetzt werden können, um kausale Gesetzmäßigkeiten zu identifizieren und Aussagen über Ursachen-Wirkungszusammenhänge zu treffen. Dies ist von zentraler Bedeutung für das experimentelle Wissen aller empirischen Wissenschaften.
- Die temporale Logik macht Aussagen über die Zeitverhältnisse von Ereignissen, wobei es nicht nur um einzelne Zeitpunkte, sondern oft um Dauern geht.

### 9.11.1 Inferenzanalyse

Zur Inferenzanalyse verwenden wir derzeit hauptsächlich die Bool'sche Logik, die wir gerade kennengelernt haben. Daneben gibt es noch den Gänzenkalkül für Satzableitungen und kausale Regularitäten, die jedoch noch nicht wirklich funktionsfähig sind, aber in naher Zukunft entwickelt werden.

## 9.12 Epistemische Inferenz und Hermeneutik

Die epistemische Inferenz ist ein Bereich, der von der theoretischen Philosophie behandelt wird und für die Leistungsfähigkeit von AI-Modellen von besonderem Interesse ist. Hier geht es darum, Wahrheitsgründe aufgrund von Thesen anzugeben, wie wir es bei der Argumentanalyse zum Klimawandel gesehen haben. Es geht um Fragen wie:

- Welche Wahrheitsgründe liegen für bestimmte Ansprüche in der Wissenschaft vor?
- Welche Gründe sind umstritten oder werden bestritten?
- Warum sind Sachbefunde korrekt oder wahr?
- Was sind die rechtfertigenden Gründe?

All dies gehört zur epistemischen Evidenz und es gibt klare Verfahren, um diese zu identifizieren, wie wir bei der Argumentanalyse gesehen haben. Dazu gehören auch die Konsistenzprüfung auf interne Widersprüche und die Wissenskritik, bei der geprüft wird, ob beispielsweise die Rechtfertigungsgründe für eine These einer Kritik standhalten oder ob Lücken identifiziert werden können, die durch weitere Evidenz oder Studien geschlossen werden müssen.

Zur epistemischen Inferenz zähle ich aber auch die Hermeneutik, die sich damit befasst, unter welchen Bedingungen und mit welchen Gründen eine bestimmte Verständnisweise von vorliegenden Texten gerechtfertigt werden kann. Hier geht es um das Verstehen von Texten und Inhalten, also genau das, wozu KI in der Lage ist: Bedeutungen zu identifizieren. Dies geschieht nicht einfach so, sondern folgt bestimmten Regeln - den hermeneutischen Regeln des Bedeutungserschließens. Diese geben vor, wie man prüfen und dazu gelangen kann, die Inhalte von symbolischen Systemen zu erschließen.

### 9.12.1 Erschließung alter Sprachen

Ein faszinierendes Beispiel für die Anwendung hermeneutischer Regeln ist die Erschließung alter Sprachen wie der Hieroglyphen im alten Ägypten. Glücklicherweise fand man hier einen Übersetzungsstein mit einem hieroglyphischen Text, einer griechischen Übersetzung und einer koptischen Version, was ausreichte, um daraus ein System für diese Sprachen zu entwickeln. Es gibt jedoch immer noch Textzeugen, von denen man zwar annimmt, dass sie Sprachen enthalten, aber nicht weiß, um welche Sprache mit welchem Inhalt es sich handelt. Auch diese können mit hermeneutischen Verfahren bearbeitet und möglicherweise entschlüsselt werden.

### **9.13 Analyse der Wissenschaftsliteratur**

Ein weiterer wichtiger Kompetenzbereich für zukünftige KI-Engines ist die Analyse der wissenschaftlichen Literatur. Wenn es stimmt, dass in der global zugänglichen Forschungsliteratur das gesamte Wissen der Menschheit enthalten ist, egal zu welcher Zeit und an welchem Ort auf unserem Globus, dann besteht die große Aufgabe der KI darin, die Inhalte dieser Literatur zu erschließen. Bisher gibt es dies noch nicht einmal im Ansatz. Digitalisierte Bibliotheken beschränken sich meist auf die Erschließung der Bibliothekskataloge, sodass man zwar feststellen kann, ob ein bestimmtes Werk in der Bibliothek vorhanden ist, und im besten Fall auf das PDF zugreifen kann, aber die eigentlichen Inhalte werden nicht erfasst. Der nächste Schritt wird also darin bestehen, die Inhalte des gesamten Wissens der Menschheit zu erfassen.

### **9.14 Das Handlungsmodell**

Das Handlungsmodell, das wir heute kennengelernt haben, ist dasjenige, das alle KI-Modelle umsetzen, jedoch mit der Modifikation, dass sie eine Instruktion zur Kenntnis nehmen. In Zukunft werden wir wohl nicht nur Fotos von Instruktionen haben, sondern ganze Instruktionsbibliotheken, die für bestimmte Problemstellungen Verfahren angeben, wie damit umzugehen ist.

In einer Chatumgebung könnten die Instruktionen dann so reformuliert werden, dass man nachvollziehen kann, ob die Umsetzung durch die Maschine die Absichten der Anfrage erfüllt. Zudem sollte es eine Erklärung aller Ausführungsschritte geben, die zugleich auch eine Erklärung des Befundes und des Ergebnisses darstellt. Dies wäre die Normalform, in der jede Handlungsausführung eines solchen Modells gesammelt und ausgewertet werden sollte.

# 10 Validierung

## 10.1 Einleitung und Rückblick

Guten Tag, meine Damen und Herren! Ich begrüße Sie herzlich zur bereits neunten Vorlesung unseres spannenden Kurses. In einem rasanten Tempo haben wir bereits mehr als die Hälfte der Themen bearbeitet und befinden uns auf der Zielgeraden.

In der letzten Stunde haben wir den AI-Akteur als zentrale Komponente zur Organisation der Aktivitäten von KI-Modellen eingeführt. Wir charakterisierten seine Fähigkeiten und erkannten, dass für ein umfassendes Verständnis der heutigen AI noch einige Puzzleteile fehlen.

### 10.1.1 Kompetenzen der AI-Akteure

Im Projekt Magister Faustus identifizierten wir bereits die Kernkompetenzen der AI-Akteure:

- Sprachkenntnisse
- Erweiterbare Fähigkeiten
- Beeinflussbarer Stil und Charakter

Den Akteuren liegen konzeptionell philosophische Theorien des rationalen Handelns zugrunde, die wir letzte Stunde in den Grundzügen ansprachen.

### 10.1.2 Erweiterungen der Kernkompetenzen

Doch diese Kernkompetenzen bedürfen noch einiger Erweiterungen, um das volle Potenzial der AI auszuschöpfen:

- **Semantik:** Die Verarbeitung von Bedeutungen bildet den revolutionären Kern, ist jedoch noch nicht vollständig implementiert.

- **Inferenzschlüsse:** Aussagenlogische und andere Schlussfolgerungen wurden bisher nur ansatzweise studiert und eingebaut.
- **Kausalität:** Unser Wissen über Ursache-Wirkungs-Beziehungen in Natur und Umwelt fehlt den Modellen noch gänzlich.
- **Epistemologie:** Auch der Bereich des Wissens und der Wissenskritik wird erst am Ende der Vorlesung thematisiert werden.

#### **10.1.2.1 Problemlösungsvermögen und Methodenwissen**

Teilweise eingebaut und berücksichtigt ist hingegen das Problemlösungsvermögen - die Fähigkeit, eine allgemeine Aufgabe in lösbare Teilschritte zu zerlegen.

Methodenwissen zur Steuerung von Instrumenten und Robotern wird intensiv beforscht, ist aber noch nicht serienreif.

Mit diesen Erweiterungen scheinen wir fast alle Erfordernisse zusammen zu haben, um das zu erschaffen, was wir als AI oder KI bezeichnen. Doch ein entscheidender Aspekt fehlt noch.

## **10.2 User Interface - Die Schnittstelle zwischen Mensch und Maschine**

Bisher nur implizit angesprochen, spielt die Interaktion zwischen AI-Modellen und uns Nutzern eine zentrale Rolle. Diese Schnittstelle, in der Informatik oft "User Interface" (UI) genannt, befindet sich derzeit im Wandel.

### **10.2.1 Expansion der Sprachsteuerung**

Die Interaktion über Sprache expandiert massiv. Instruktionen können direkt über Mikrofone, Ohrhörer und Audiogeräte übergeben werden. Sprachassistenten werden in Textverarbeitungsprogramme integriert und übernehmen weit mehr als nur die Rechtschreibkontrolle.

### **10.2.2 Aufstieg der AI-Bots**

Noch wenig bekannt, aber mit enormem Zukunftspotenzial sind AI-Bots - selbstständig arbeitende Roboter. Wir kennen sie bereits von automatischen Backups oder E-Mail-Checks. Bis Ende des Jahres

werden sie massiv ausgebaut und in vielen Bereichen wie Smart Homes eingesetzt werden können.

## 10.3 Bedeutung von Wissensquellen

Ein oft übersehener, aber extrem wichtiger Faktor für die Funktionalität der AI sind die verwendeten Wissensquellen und Informationsressourcen.

- **Wikipedia:** Die Wikipedia-Artikel in allen Sprachen bilden den Kernbestand des Hintergrundwissens der meisten Sprachmodelle.
- **Trainingsdaten:** Der Großteil sind jedoch die trainierten Sprachdaten, die durch die Interaktion mit Nutzern ständig erweitert werden und so auch aktuelle Themen umfassen.
- **Zeithorizont:** Die Modelle haben derzeit den Diskussionsstand im Internet vom November letzten Jahres. Alles bis dahin diskutierte ist als Wissen verfügbar.## Mehr als nur natürliche Sprache: KI-Modelle als Programmierer

Es ist faszinierend zu beobachten, wie KI-Modelle nicht nur in der Lage sind, natürlichsprachliche Sätze zu formulieren, sondern auch kompetente Programme in vielen verschiedenen Programmiersprachen zu entwickeln. Dieses beeindruckende Können wird jedoch selten diskutiert, obwohl es die Grundlage für viele dieser Modelle bildet. Wie ist es möglich, dass die Sprachmodelle über solch ein umfassendes Programmier-Know-how verfügen?

### 10.3.1 GitHub: Eine Schatzkammer für KI-Modelle

Der Schlüssel zu diesem verborgenen Wissen liegt in der Entwicklungsgeschichte der KI-Modelle. Die Entwickler haben zwei große Programm- oder Textdatenbestände ausgewertet, die entscheidend für den Erfolg waren. Einer davon ist GitHub, ein Tool zum Speichern und Austauschen von Programmcode, das vor etwa fünf oder sechs Jahren von Microsoft aufgekauft wurde. Vor der Übernahme war GitHub ein kostenloses Open-Access-Tool, das von vielen Programmierern genutzt wurde, um ihren Code sicher zu speichern, auszutauschen und mit einer Revisionsgeschichte zu verwalten.

Microsoft erwarb den gesamten Datenbestand von GitHub einschließlich der Nutzungsrechte. Dieser Teil bildet den Hintergrund für die Modelle von OpenAI und fast allen anderen bekannten KI-Systemen. Durch das Training an allen weltweit erstellten Programmen in diesen Dateien haben die KI-Modelle gelernt, wie Prozeduren und Algorithmen in bestimmten Programmiersprachen dargestellt werden. Dieses Wissen wurde vollständig erfasst.

### **10.3.2 Von Fehlerdiskussionen zu Syntax-Änderungen**

Neben GitHub gibt es noch weitere Diskussionsforen, in denen Programmierer über Jahre hinweg diskutiert haben, wie mit bestimmten Fehlern in Programmen umzugehen ist und wie diese gelöst werden können. Diese Diskussionsverläufe sind in den Trainingsdaten rekonstruierbar und geben Aufschluss über die Weiterentwicklung von Programmen. Anhand von KI-generierten Programmvorschlügen lässt sich sogar erkennen, zu welchem Zeitpunkt der Entwicklung diese Programmbibliotheken konsultiert und eingebaut wurden. Beispielsweise können neue Versions-Syntax-Definitionen dazu führen, dass Programme nicht mehr rückwärtskompatibel sind.

### **10.3.3 Die Macht der Programmiersprachenkenntnisse**

Die im Hintergrund verborgene Wissensquelle der KI-Modelle ist also eine umfassende Programmiersprachenkenntnis, die fast alles umfasst, was jemals im öffentlichen Datenraum programmiert wurde. Diese Quelle wird auch heute noch genutzt und verleiht den KI-Systemen ihre beeindruckenden Fähigkeiten.

## **10.4 Neue Informationsquellen für KI-Modelle**

### **10.4.1 Websuche: Aktuelles Wissen ergänzt historische Daten**

Lange Zeit wurde die Websuche von KI-Modellen eher stiefmütterlich behandelt, da es hauptsächlich darum ging, bereits existierende Webseiten wiederzufinden und in neuer Formatierung anzuzeigen. Doch seit einem halben Jahr werden diese Daten vermehrt konsultiert, da sie aktuelle Informationen liefern, die über die historischen Datenbestände hinausgehen. Zusätzlich können Web-Funde dazu beitragen, sogenannte "Halluzinationen" - sprachlich wohlgeformte, aber inhaltlich erfundene Informationen - zu kontrollieren und auf ihre Richtigkeit zu überprüfen.

### **10.4.2 Informationsbroker: Spezialisierte und verlässliche Datenquellen**

In Zukunft werden Informationsbroker eine entscheidende Rolle spielen. Dabei handelt es sich um spezialisierte Informationsquellen, die für bestimmte Datenbereiche verlässliche Informationen bereitstellen. Dazu gehören beispielsweise staatliche Informationen wie Steueraufkommen, Einwohnermeldezahlen, Gesetzestexte und viele weitere Daten, an deren korrekter Wiedergabe der Staat und

die Europäische Union interessiert sind. Seit etwa zwei Monaten werden diese Informationsbroker von KI-Modellen genutzt, was zuvor noch nicht der Fall war.

## 10.5 Herausforderungen und Grenzen von KI-Modellen

### 10.5.1 Rechtssammlung: Ein Beispiel für Interpretationsschwierigkeiten

Lassen Sie uns anhand eines Beispiels verdeutlichen, wie KI-Modelle mit Informationsquellen umgehen und welche Herausforderungen dabei auftreten können. Betrachten wir dazu die Rechtssammlung, insbesondere das Urhebergesetz. Überall in Europa sind Gesetzestexte öffentlich zugänglich und können im Internet aufgerufen werden. Dort finden sich die geltenden Regeln zum Umgang mit geschützten Werken und den Rechten des Urhebers.

Eine typische Frage in diesem Zusammenhang könnte lauten: “Ist mein AI-Programm ein geschütztes Werk?” Um diese Frage zu beantworten, können wir das aktuelle Apple-User-Interface nutzen, das es erlaubt, mit einem einfachen Tastendruck ein Fenster aufzurufen und die Frage dort einzugeben. Das Programm fotografiert dann die geöffnete Seite, transkribiert sie in Text und wertet diesen mit Sprachverarbeitung aus.

Die generierte Antwort ist oft länger als der relevante Ausschnitt des Gesetzestextes und vermischt verschiedene Informationen wie Patentrecht, Markenrecht und Zusammenfassungen. Es wird versucht, eine Antwort zu formulieren, die sich auf die Frage bezieht, ob Computerprogramme urheberrechtlich geschützt sein können. Allerdings enthält die Antwort oft ungenaue oder sogar falsche Formulierungen, die rechtlich keinen Sinn ergeben.

### 10.5.2 Halluzinationen und fehlende kritische Hinterfragung

Das Problem der “Halluzinationen” - sprachlich sauber formulierte, aber inhaltlich erfundene Informationen - tritt auch bei der Rechtsinterpretation auf. Die generierten Antworten klingen überzeugend und vollständig, sind aber nicht immer korrekt. Es fehlt die Berücksichtigung einschlägiger Urteile von oberen Gerichten, die zwar über Rechtssammlungen verfügbar wären, aber bisher nicht einbezogen werden.

Was vollständig fehlt, ist die kritische Hinterfragung und methodische Bewertung der generierten Auskünfte. Die Programme können zwar mit normativen Formulierungen wie in Gesetzen umgehen, haben aber keine verfahrenskritische Kompetenz, um zu überprüfen, auf welche rechtlichen Normen sich eine Anfrage stützt, ob diese noch gültig sind und wie sie mit der Rechtspraxis vereinbar sind.

### **10.5.3 Mögliche negative Folgen und Regulierungsbedarf**

Diese Schwächen der KI-Modelle können extreme negative Folgen nach sich ziehen und müssen mit Sicherheit geregelt werden. Es ist faszinierend, dass die Programme Informationen aus dem Internet abrufen, verarbeiten und daraus Antworten generieren können. Doch über den verantwortungsvollen Umgang mit diesen Fähigkeiten wurde bisher zu wenig nachgedacht.

## **10.6 Ausblick: Weitere Anwendungsbereiche und RAG**

### **10.6.1 Informationsbroker in verschiedenen Bereichen**

Informationsbroker werden nicht nur im Rechtsbereich eine wichtige Rolle spielen, sondern auch in anderen Gebieten wie Freizeit und Tourismus. Wenn es um die Suche nach günstigen Flügen oder Ferientielen geht, müssen Nutzer derzeit noch manuell im Netz recherchieren und entsprechende Informationsbroker nutzen. Es ist jedoch nur eine Frage der Zeit, bis auch diese Informationen für KI-Modelle zugänglich gemacht werden.

Auch Kulturgüter und Sammlungsbestände von Museen und Archiven werden in Zukunft verstärkt in die Sphäre der KI eingehen. Bisher werden diese Informationen hauptsächlich über die Webseiten der jeweiligen Institutionen zur Verfügung gestellt, aber noch nicht von KI-Systemen genutzt.

### **10.6.2 Retrieval Augmented Generation (RAG)**

Ein vielversprechender Ansatz, um die Schwächen der KI-Modelle zu überwinden, ist die sogenannte Retrieval Augmented Generation (RAG). Dabei werden zusätzliche Informationsquellen genutzt, um fehlende oder vorgetäuschte Informationen zu ergänzen und zu korrigieren.

Der Prozess besteht aus drei Schritten:

1. Retrieval: Relevante Informationen werden aus externen Quellen gesucht und gefunden.
2. Augmentation: Die gefundenen Informationen werden aufbereitet, komprimiert und in die Anfrage integriert.
3. Generation: Aus den angereicherten Informationen wird eine Antwort generiert.

Dieser Ansatz wurde beispielsweise bei der Beantwortung der Frage zum Urheberrecht angewendet, indem der Gesetzestext als zusätzliche Informationsquelle herangezogen wurde.## Nutzung externer Wissensquellen durch AI-Modelle

In der heutigen Vorlesung möchte ich Ihnen einen faszinierenden Aspekt der Weiterentwicklung von AI-Systemen näherbringen: Die Nutzung externer Wissensquellen, um die Antwortqualität und Verlässlichkeit der Modelle zu verbessern. Lassen Sie uns gemeinsam ergründen, welche Möglichkeiten sich hier eröffnen und wie wir diese in Zukunft optimal ausschöpfen können.

### **10.6.3 Die RAG-Komponente**

Stellen Sie sich vor, Sie spezifizieren eine externe Quelle, beispielsweise die deutschen Gesetzestexte zum Urheberrecht, binden diese in die Anfrage eines Nutzers ein und erwarten vom Programm eine fundierte Aufbereitung der Antwort. Genau hier kommt die RAG-Komponente ins Spiel. An dieser vielversprechenden Technologie wird derzeit intensiv geforscht und entwickelt. Mittlerweile ist es möglich, selbst sehr große PDF-Sammlungen effizient zu indizieren. Spezielle Programme filtern dann die relevanten Seiten aus Hunderttausenden von PDF-Dokumenten heraus. Im Optimalfall werden diese herausgefilterten Stellen noch weiter komprimiert, sodass sie nahtlos in die Anfrage integriert werden können, die man dem KI-Modell für eine präzise Antwort übergibt. Wissensdatenbanken gewinnen in diesem Kontext zunehmend an Bedeutung.

### **10.6.4 Die Dimensions-Datenbank**

Lassen Sie mich Ihnen nun eine beeindruckende Wissensdatenbank vorstellen, die den riesigen, weltweit verfügbaren Wissensbestand exemplarisch veranschaulicht und für zukünftige AI-Modelle nutzbar macht: Dimensions. Auch wenn die Integration solcher Datenbanken momentan noch nicht standardmäßig vorgesehen ist, bin ich überzeugt, dass sich dies in absehbarer Zeit ändern wird.

Ein Kernproblem, an dem viele große Provider derzeit arbeiten, ist das Halluzinieren von AI-Modellen. Eine mögliche Lösung besteht darin, nicht nur Wikipedia als Quelle heranzuziehen, sondern auch wissenschaftliche Preprint-Server wie arXiv auszuwerten. Genau hier kommt Dimensions ins Spiel - eine englischsprachige, nutzbare Datenbank, auf die wir über die HU-Uni zugreifen können.

#### **10.6.4.1 Wissenschaftliches Publikationswesen im Wandel**

Die Zugänglichkeit wissenschaftlichen Wissens hat sich in den letzten Jahren radikal verändert, insbesondere durch die Corona-Pandemie. Zuvor waren viele Publikationen hinter Bezahlschranken der Verlage verborgen und erst Monate später öffentlich zugänglich. Doch mit Ausbruch von Corona wurden diese

Schranken praktisch vollständig aufgehoben. Verlage machten Publikationen bereits vor dem Review-Prozess frei verfügbar. Dies führte zu einer erheblich gesteigerten Wissensdynamik. Über Datenbanken wie Dimensions konnte man nun tagesaktuell den Publikationsstand zu beliebigen Themen abfragen. Innerhalb des ersten Jahres der Pandemie waren so bereits 600.000 Publikationen zu Corona in all seinen Facetten zugänglich und auswertbar.

### **10.6.5 Der digitalisierte Wissensschatz**

Um zu verdeutlichen, auf welche umfangreiche Informationen sich AI-Modelle zukünftig stützen können, habe ich eine Abfrage in Dimensions durchgeführt. Das Ergebnis ist beeindruckend: Der Gesamtbestand des digitalisierten wissenschaftlichen Wissens ist keineswegs ein ferner Zukunftstraum. In vielen Disziplinen, insbesondere im naturwissenschaftlich-medizinisch-biologischen Bereich, ist er bereits in hohem Maße vorhanden und zugänglich.

Geisteswissenschaften und Literaturwissenschaften hinken hier allerdings noch hinterher. Bücher sind bis heute kaum online verfügbar und auswertbar - trotz mehrerer geförderter EU-Programme, die dies ändern sollten.

#### **10.6.5.1 Beeindruckende Publikationszahlen**

Lassen Sie uns einen genaueren Blick auf die Suchergebnisse in Dimensions werfen. Zu den Stichworten "Large Language Models and AI" finden sich sage und schreibe 1,2 Millionen Publikationen. Ein Blick auf den Kurvenverlauf der Häufigkeiten über die Jahre offenbart Erstaunliches: Entgegen meiner Vermutung, dass Large Language Models erst seit zwei Jahren intensiv erforscht werden, reichen die Publikationen viel weiter zurück.

Für bestimmte Themenbereiche wie künstliche Intelligenz und Machine Learning finden sich jeweils um die 100.000 Publikationen. Doch auch in den Naturwissenschaften (Physical Sciences) und Biowissenschaften (Biological Sciences) gibt es bereits Zehntausende von Veröffentlichungen zu diesen Themen. Ab 2015 sind jährlich 3.600 Publikationen zu verzeichnen - eine beachtliche Zahl, die ich so nicht erwartet hätte. Und der Trend geht in einer charakteristischen exponentiellen Kurve steil nach oben, über viele Disziplinen und Themenfelder hinweg.

### **10.6.5.2 Scientometrische Aufschlüsselung**

Die Datenbank bietet nicht nur Zugang zu den Publikationen selbst, sondern auch eine detaillierte scientometrische Aufschlüsselung. Autorennamen, Institutionen, Finanziers, Verlage, Open-Access-Status - all diese Metadaten wurden sorgfältig erfasst und konsistent aufbereitet. Zu jeder der 100 Millionen wissenschaftlichen Publikationen existiert mindestens ein Abstract, das den Inhalt zusammenfasst. Die meisten Volltexte, insbesondere im Open-Access-Bereich, sind zudem als PDF öffentlich zugänglich.

Diese granulare Erschließung ermöglicht es, die Wissensproduktion der letzten 30 Jahre sehr genau zurückzuverfolgen. Auch ältere Publikationen, etwa von Einstein, sind verzeichnet, wenn auch lückenhafter. Die große Herausforderung für die AI besteht nun darin, diesen gigantischen Wissensbestand effektiv zu nutzen - eine Aufgabe, die bisher noch nicht zufriedenstellend gelöst ist.

### **10.6.6 Zugang zu wissenschaftlichen Publikationen**

Um den Zugang zu wissenschaftlichen Publikationen zu erleichtern, haben sich Deutschland, Österreich und vermutlich auch die Schweiz zu einem Gesamtforschungskonsortium zusammengeschlossen. Ziel ist es, mit den großen Publikationshäusern Verträge abzuschließen, die deren gesamtes Programm öffentlich zugänglich machen. Diese Vereinbarungen kosten zwar Millionen, ermöglichen aber einen unschätzbaren Mehrwert für die Forschungsgemeinschaft.

Die Verlagslandschaft hat sich in den letzten Jahren stark konsolidiert. Weltweit gibt es nur noch eine Handvoll großer wissenschaftlicher Verlage, die ganze Themenfelder mit ihrem Programm abdecken. Durch die Konsortialverträge können nun Universitäten wie die Humboldt-Universität auf das gesamte Angebot dieser Verlage zugreifen und Artikel direkt herunterladen - ein enormer Fortschritt gegenüber früheren Bestellprozessen über Bibliotheken.

### **10.6.7 Die Wahrheitsfrage**

Doch mit dem Zugang zu dieser Fülle an Informationen stellt sich zwangsläufig die Frage nach deren epistemischer Bewertung. Stimmen die Aussagen in den Publikationen? Worauf gründet sich ihre Wahrheit? Diese Frage soll in Zukunft für jeden Satz, jedes Diagramm einer beliebigen Veröffentlichung beantwortet werden können - eine Mammutaufgabe für die AI.

Nehmen wir als Beispiel den folgenden Satz aus einem renommierten Journal: "Im Jahr 2020 wurde GPT-3 veröffentlicht mit 175 Billionen, also Milliarden Parametern, das selbst 100 Mal größer war als der Vor-

läufer GPT-2.” Wie können wir nun mithilfe von AI-Tools überprüfen, ob diese Aussage wahr ist? Lassen Sie es uns gemeinsam ausprobieren!

#### **10.6.7.1 Ein Test mit ChatGPT**

Ich markiere den relevanten Satz und öffne über die Tastenkombination “Option-Leertaste” den Zugang zu ChatGPT. Dieses Tool soll mir nun, bezogen auf die Markierung, Auskunft geben, ob die Information korrekt ist. In Zukunft erwarte ich, dass dies für jegliche Inhalte im Internet möglich sein wird - seien es historische Quellen, Politikeräußerungen oder wissenschaftliche Publikationen. Die Wahrheitsfrage ist dabei von zentraler Bedeutung.

Wie Sie sehen, ist der gesamte Apparat bereits darauf ausgelegt, beliebige dieser Hunderte von Millionen Publikationen zu analysieren. Es bleibt spannend zu beobachten, wie sich die AI-Modelle dieser Herausforderung stellen werden. Eines ist sicher: Wir stehen hier erst am Anfang einer faszinierenden Entwicklung, die unser Verständnis von Wissen und Wahrheit nachhaltig prägen wird.## Einführung in die Überprüfung von Aussagen mithilfe von AI

Zu Beginn der heutigen Vorlesung möchte ich Ihnen eine spannende neue Funktion vorstellen, die es ermöglicht, die Wahrheit von Aussagen mithilfe von AI zu überprüfen. Obwohl die Live-Vorführung immer mit gewissen Tücken verbunden ist, lassen Sie uns gemeinsam erkunden, was dieses System zu leisten vermag.

#### **10.6.8 Überprüfung einer Aussage zu GPT-3 Parametern**

Zunächst habe ich dem System einen Screenshot mit einer markierten Aussage zu den Parameterzahlen von GPT-2 und GPT-3 bereitgestellt. Die AI hat den markierten Satz korrekt identifiziert und überprüft, ob dieser wahr ist. Dazu wurden Quellen herangezogen - ein offizieller OpenAI-Blog und ein Preprint-Artikel. Vor einem halben Jahr wäre eine solche Leistung undenkbar gewesen!

Natürlich kann man über die Qualität der Quellen diskutieren, insbesondere wenn es sich um nicht peer-reviewte Preprints handelt. Dennoch ist es beeindruckend, dass überhaupt passende Quellen gefunden und referenziert wurden.

### **10.6.9 Grenzen der AI bei philosophischen und geisteswissenschaftlichen Anfragen**

Es ist wichtig zu beachten, dass diese Überprüfung nicht bei allen Arten von Aussagen gleich gut funktioniert. Insbesondere bei philosophischen oder geisteswissenschaftlichen Themen stößt die AI an ihre Grenzen. Hier fehlt es häufig an klar definierten Wahrheitskriterien und eindeutigen Quellen.

## **10.7 Bewertung der Leistungsfähigkeit und Grenzen der AI**

### **10.7.1 Imitation einer Rechtfertigung statt echter Begründung**

Obwohl die Ergebnisse auf den ersten Blick beeindruckend wirken, dürfen wir nicht vergessen, dass es sich hierbei lediglich um eine Imitation einer Rechtfertigung handelt. Die AI ist nicht in der Lage, die Argumentation in den zitierten Publikationen wirklich zu prüfen und zu bewerten. Es fehlt an echten Rechtfertigungsmodellen, die eine unabhängige Beurteilung der Evidenz ermöglichen würden.

### **10.7.2 Fehlende Berücksichtigung von Gegenargumenten und alternativen Sichtweisen**

Ein weiteres Defizit besteht darin, dass die AI keine Publikationen berücksichtigt, die der überprüften Aussage widersprechen könnten. Auch alternative Beurteilungen der Wahrheit einer Aussage werden nicht einbezogen. Für eine umfassende epistemische Bewertung wären jedoch genau solche Aspekte von großer Bedeutung.

### **10.7.3 Potenzial für die Untersuchung kontroverser Studien**

Interessante Anwendungsmöglichkeiten sehe ich in der Untersuchung kontroverser Studien, wie beispielsweise der Behauptung supraleitender Keramiken bei Zimmertemperatur im letzten Jahr. Hier konnte die AI immerhin ähnliche kritische Anforderungen identifizieren wie die Peer-Reviewer. Das lässt hoffen, dass in Zukunft zumindest eine grobe Einschätzung der wissenschaftlichen Qualität von Publikationen möglich sein wird.

## 10.8 Ausblick auf zukünftige Entwicklungen

### 10.8.1 Anforderungen an eine epistemische Bewertung durch AI

Um eine wirklich zuverlässige Überprüfung von Aussagen durch AI zu ermöglichen, müssen wir an folgenden Punkten arbeiten:

- Themenfokussierung bis auf Satzebene unter Berücksichtigung von Fachsprachen und mathematischen Verfahren
- Inferenzielles Retrieval zur Überprüfung von Rechtfertigungen und Begründungen
- Beantwortung von Fragen nach Konsequenzen, Widersprüchen und Zusammenhängen zu Vergleichspublikationen
- Kombination von Text- und Bildanalyse zur Aufdeckung von Manipulationen

Erst wenn wir diese epistemische Bewertung der zugrundeliegenden Datenbestände gemeistert haben, können wir ernsthaft daran denken, Fake News von korrekten Informationen zu unterscheiden.

### 10.8.2 Bausteine für die zukünftige Forschung

In den kommenden Vorlesungen möchte ich Ihnen einige Bausteine vorstellen, die meiner Meinung nach für die Weiterentwicklung der epistemischen Bewertung durch AI von Bedeutung sind. Insbesondere die Analyse wissenschaftlicher Texte hinsichtlich ihrer Gehalte und die Lösung kausaler Hypothesen werden dabei eine wichtige Rolle spielen.

Poppers Behauptung, dass es sich bei kausalen Redeweisen um Konditionale handelt, halte ich für Unsinn. Solange wir versuchen, diesen Unsinn in die AI-Modelle einzubauen, werden wir keine vernünftigen Ergebnisse in der Bewertung erzielen können.

Mit diesen spannenden Themen werden wir uns also in der nächsten Woche intensiver beschäftigen. Bis dahin wünsche ich Ihnen eine schöne Woche und hoffe, dass Sie noch etwas Sonne tanken können!

# 11 Epistemische Bewertung

## 11.1 Die philosophische Vogelperspektive auf AI-Modelle

Lassen Sie uns nun tiefer in die Materie eintauchen und die Folgerungen betrachten, die sich aus unserer philosophischen Vogelperspektive ergeben. Besonders interessant ist dabei der Aspekt, mit dem ich die letzte Vorlesung abschloss: die epistemische Bewertung dessen, was AI-Modelle erzeugen und generieren.

### 11.1.1 Dimensionen der epistemischen Bewertung

Epistemische Bewertungen umfassen verschiedene Arten von Urteilen:

- Die Beurteilung von Aussagen als wahr oder falsch
- Die Kritik an Behauptungen anderer über Wahrheitswerte von Aussagen
- Die Rechtfertigung durch Angabe von Gründen, warum eine Aussage wahr oder falsch ist

Diese Dimension der epistemischen Bewertung ist erst jetzt in den Fokus gerückt. Früher hatten wir es lediglich mit Imitationen oder Simulationen zu tun - eine erste Annäherung. Doch der eigentliche Zugang zu dieser Dimension eröffnet sich erst mit der gegenwärtigen AI-Revolution.

### 11.1.2 Ein herausforderndes Puzzle für AI-Modelle

Um dies zu veranschaulichen, möchte ich ein Beispiel heranziehen, das in der AI-Szene vor etwa vier Wochen für Aufsehen sorgte. Es handelt sich um folgendes Puzzle:

“Egon hat drei Brüder. Jeder Bruder hat zwei Schwestern. Wie viele Schwestern hat Egon?”

Interessanterweise konnten die meisten AI-Modelle dieses scheinbar einfache Rätsel nicht lösen. Doch bevor wir uns die Antworten der Modelle ansehen, ist es wichtig zu verstehen, dass die großen AI-

Entwickler nach Bekanntwerden dieses Problems rasch daran arbeiteten, ihren Modellen diese Art von Puzzles beizubringen.

## **11.2 Analyse der AI-Antworten**

Lassen Sie uns nun betrachten, wie verschiedene AI-Modelle auf diese Herausforderung reagieren. Wir beginnen mit einem der schnellsten und leistungsfähigsten Modelle, Anthropic Cloud Haiku.

### **11.2.1 Anthropic Cloud Haiku's Antwort**

Das Modell antwortet:

“Okay, lasst uns das Schritt für Schritt durchrechnen. Egon hat drei Brüder, jeder Bruder hat zwei Schwestern. Das bedeutet, dass Egon insgesamt dreimal zwei, sechs Schwestern hat.”

Diese Antwort offenbart ein fundamentales Problem: Sie klingt überzeugend, ist aber falsch. Der Kern des Problems liegt darin, dass man an der Antwort selbst nicht erkennen kann, ob sie richtig oder falsch ist, es sei denn, man rechnet selbst nach und kontrolliert das Ergebnis.

### **11.2.2 Das Kernproblem der epistemischen Dimension**

Das eigentliche Problem ist nicht, dass AI-Modelle gelegentlich Fehler machen. Vielmehr liegt es darin, dass sie ihre Antworten mit dem Brustton der Überzeugung präsentieren, unabhängig davon, ob diese korrekt sind oder nicht. Dies kann in verschiedenen Kontexten, von der Rentenversicherung bis hin zu anderen Bereichen, schwerwiegende Folgen haben.

## **11.3 Die Grenzen aktueller AI-Modelle**

Es ist wichtig zu verstehen, dass kein gegenwärtiges AI-Modell die epistemische Dimension vollständig erfassen und lösen kann. Die aktuellen Modelle sind ausschließlich darauf trainiert, sprachlich wohlgeformte Sätze zu generieren. Die epistemische Dimension, obwohl sie mit Aussagen zu tun hat, ist noch nicht implementiert.

### **11.3.1 Methodische Fehler im Reasoning-Apparat**

Die entscheidende Frage lautet: Was geht methodisch falsch im Reasoning-Apparat, dem Nachdenkmechanismus der AI-Modelle? Es handelt sich hier nicht um logische Schlüsse im klassischen Sinne. Das Problem liegt tiefer und erfordert eine gründlichere Analyse.

## **11.4 Verbesserungen und neue Ansätze**

Nach dem initialen Scheitern bei diesem Puzzle haben die größeren Modelle begonnen, ihre Herangehensweise zu überdenken. Lassen Sie uns nun betrachten, wie neuere Modelle, insbesondere Claude Sonnet von Anthropic, das derzeit als Spitzenreiter in der Qualität des Reasonings gilt, mit dieser Aufgabe umgehen.

### **11.4.1 Claude Sonnet's Lösungsansatz**

Claude Sonnet gibt folgende Antwort:

“Der Schlüssel zur Lösung liegt darin zu erkennen, dass Egon und seine Brüder die gleichen Schwestern haben. Wenn jeder Bruder zwei Schwestern hat, dann hat auch Egon zwei Schwestern. Die Anzahl der Brüder spielt für die Anzahl der Schwestern keine Rolle. Daher die Antwort: Egon hat zwei Schwestern.”

Diese Antwort ist korrekt und zeigt, dass die Entwickler an genau der Stelle nachgebessert haben, wo wir zuvor das Problem identifiziert hatten.

### **11.4.2 Vergleich mit anderen Modellen**

Um einen umfassenderen Überblick zu erhalten, lassen Sie uns auch die Antwort von ChatGPT betrachten:

“Egon hat zwei Schwestern. Egon hat drei Brüder, jeder Bruder hat zwei Schwestern. Daraus folgt, dass Egon selbst einer der Brüder ist. Da jeder Bruder zwei Schwestern hat und Egon einer dieser Brüder ist, hat er ebenfalls zwei Schwestern.”

Diese Antwort ist ebenfalls korrekt, wenn auch etwas knapper in der Begründung.

## 11.5 Die Herausforderung der epistemischen Sicherheit

Trotz der korrekten Antworten bleibt eine fundamentale Frage bestehen: Wie können wir sicher sein, dass die Antwort stimmt? Diese Unsicherheit ist ein zentrales Problem im Umgang mit AI-Modellen. Es erfordert von uns, stets kritisch zu hinterfragen und zu überprüfen, selbst wenn die Antworten überzeugend klingen.

### 11.5.1 Ein komplexeres Beispiel: Goethe und das Kartenspiel

Um die Herausforderung der epistemischen Bewertung weiter zu verdeutlichen, betrachten wir ein komplexeres Beispiel aus der Biografie von Johann Wolfgang von Goethe. Die Frage lautet: "Spielte Goethe Karten?"

Diese scheinbar einfache Frage öffnet die Tür zu einer Fülle von Informationen und Interpretationen seitens der AI-Modelle. Es zeigt, wie selbst bei vermeintlich simplen Anfragen die Modelle dazu neigen, umfangreiche und detaillierte Antworten zu generieren, die weit über die eigentliche Fragestellung hinausgehen.## Die Herausforderung der Wahrheitsfindung in der künstlichen Intelligenz

In der heutigen Vorlesung möchte ich mit Ihnen einen faszinierenden Aspekt der künstlichen Intelligenz erörtern: die Herausforderung der Wahrheitsfindung. Um dieses komplexe Thema greifbar zu machen, werden wir uns eines konkreten Beispiels bedienen - der Frage, ob Johann Wolfgang von Goethe ein begeisterter Kartenspieler war.

### 11.5.2 Die KI-generierte Antwort und ihre Schwächen

Lassen Sie uns zunächst die Antwort betrachten, die ein modernes AI-System auf diese Frage liefert. Die Ausgabe liest sich wie ein wohlformulierter Wikipedia-Artikel, der Goethe nicht nur als herausragenden Literaten, sondern auch als leidenschaftlichen Kartenspieler darstellt. Das Kartenspiel wird als integraler Bestandteil seines Lebens und seiner Persönlichkeit beschrieben, der sogar seine literarischen Leistungen beeinflusst haben soll.

Auf den ersten Blick mag diese Antwort überzeugend wirken. Doch bei näherer Betrachtung offenbaren sich gravierende Mängel. Das AI-System gibt keine Quellen an, die seine Behauptungen stützen. Es reformuliert lediglich vorhandene Informationen, ohne deren Herkunft oder Zuverlässigkeit zu hinterfragen. Dies ist ein fundamentales Problem bei der derzeitigen Implementierung von AI-Systemen.

### **11.5.3 Die Notwendigkeit einer fundierten Quellenprüfung**

Als Wissenschaftler und kritische Denker müssen wir uns fragen: Wie können wir den Wahrheitsgehalt solcher Aussagen überprüfen? Die simple Angabe von Zitaten oder das Verweisen auf andere Quellen reicht nicht aus. Wir benötigen eine rigorose Methodik zur Überprüfung historischer Fakten.

In unserem Fall wäre ein idealer Ansatz, eine von Goethe selbst autorisierte Quelle zu finden, in der er sich zum Thema Kartenspiel äußert. Dies würde uns eine solide Grundlage für die Beantwortung unserer Frage liefern.

### **11.5.4 Ein konkretes Beispiel: Goethes Brief von 1773**

Glücklicherweise verfügen wir über eine solche Quelle. Es handelt sich um einen Brief Goethes an Johann Christian Kerstner vom 14. April 1773. In diesem Schreiben erwähnt Goethe das Kartenspiel, wenn auch nicht explizit seine eigene Teilnahme daran. Er schreibt:

“Aber dass ihr, weil euch das Glück die Karten gemischt hat, mit der Spadille stecht, mir ein höllisches Gesicht zieht und euch zu eurem Weibe legt, finde ich unartig.”

Dieser Brief liefert uns wichtige Indizien. Zum einen zeigt er Goethes Vertrautheit mit der Terminologie des Kartenspiels. Zum anderen lässt der Kontext vermuten, dass Goethe selbst an solchen Spielrunden teilgenommen hat.

### **11.5.5 Die Herausforderungen bei der AI-gestützten Analyse historischer Quellen**

Die Aufbereitung solcher historischer Quellen für die Analyse durch AI-Systeme stellt uns vor beträchtliche Herausforderungen. In der Vergangenheit erforderte dies jahrzehntelange manuelle Arbeit zur Digitalisierung und Strukturierung der Texte.

Moderne AI-Modelle bieten hier neue Möglichkeiten. Sie können bereits Bildschirmfotos von Webseiten strukturell erfassen und relevante Informationen extrahieren. Dies eröffnet uns neue Wege, historische Dokumente effizienter für die AI-gestützte Analyse aufzubereiten.

### **11.5.6 Die Grenzen aktueller AI-Systeme**

Trotz dieser Fortschritte stoßen wir immer noch an die Grenzen aktueller AI-Systeme. In unserem Beispiel fragte ich das System erneut, ob Goethe im Jahr 1773 Karten spielte - obwohl wir nun über einen eindeuti-

gen Beleg verfügen.

Die Antwort des Systems war ernüchternd. Es gab zu, keine spezifischen Aufzeichnungen zu haben, spekulierte aber dennoch über Goethes mögliche Kartenspielaktivitäten basierend auf allgemeinen Praktiken der Zeit. Dies zeigt deutlich die Schwächen des Systems auf: Es hat keinen Zugriff auf die relevante Information und füllt diese Wissenslücke mit Spekulationen.

## **11.6 Ausblick: Die Zukunft der AI in der Wahrheitsfindung**

Diese Erkenntnisse verdeutlichen die immensen Herausforderungen, vor denen wir bei der Entwicklung von AI-Systemen stehen, die in der Lage sind, Wahrheitsansprüche zuverlässig zu rechtfertigen. Es ist ein komplexes Unterfangen, das nicht nur technische, sondern auch organisatorische Hürden zu überwinden hat.

Dennoch bin ich optimistisch, dass wir in absehbarer Zeit - vielleicht schon innerhalb eines Jahres - signifikante Fortschritte in dieser Richtung sehen werden. Stellen Sie sich vor, welche Möglichkeiten sich eröffnen, wenn wir über AI-Systeme verfügen, die in der Lage sind, Aussagen in Manuskripten oder politischen Debatten substanziell und unabhängig zu verifizieren.

Die Realisierung solcher Systeme wäre nichts weniger als revolutionär. Sie würde nicht nur die akademische Forschung transformieren, sondern könnte auch weitreichende Auswirkungen auf Bereiche wie die Gesetzgebung haben. Stellen Sie sich vor, Parlamentarier hätten Zugang zu einem AI-System, das komplexe Sachfragen nahezu definitiv beantworten kann - welch ein Potenzial für fundierte politische Entscheidungen!

Natürlich dürfen wir nicht der Illusion verfallen, dass solche Systeme fehlerfrei oder unfehlbar sein werden. Aber sie könnten uns ein mächtiges Werkzeug an die Hand geben, um der Wahrheit näherzukommen und besser begründete Entscheidungen zu treffen.

In diesem Sinne ist unser Projekt mit der Stiftung Weimarer Kultur, den gesamten Goethe-Nachlass für AI-Analysen aufzubereiten, ein wichtiger Schritt in diese Richtung. Es zeigt exemplarisch, wie wir historische Quellen für moderne Technologien zugänglich machen und gleichzeitig die Integrität und den Kontext dieser Quellen bewahren können.

Lassen Sie uns also weiter daran arbeiten, die Möglichkeiten der künstlichen Intelligenz zu erweitern und gleichzeitig kritisch hinterfragen, wie wir diese Technologie am besten für die Wahrheitsfindung einsetzen können. Die Herausforderungen sind groß, aber die potenziellen Gewinne für Wissenschaft, Gesellschaft und Politik sind es ebenso.## Die Herausforderungen der AI-Halluzinationen

In der faszinierenden Welt der künstlichen Intelligenz stoßen wir immer wieder auf ein Phänomen, das uns Kopfzerbrechen bereitet: die sogenannten “Halluzinationen” der AI-Systeme. Lassen Sie mich Ihnen erklären, warum dieses Problem so bedeutsam ist und welche Konsequenzen es hat.

### **11.6.1 Das Kernproblem der Falschinformationen**

Stellen Sie sich vor, Sie fragen ein AI-Programm nach einer bestimmten Information. Nun passiert Folgendes: Wenn das Programm die Information nicht findet oder nicht kennt, behauptet es kurzerhand, dass diese Information schlichtweg nicht existiert. Das ist, wie Sie sich vorstellen können, eine äußerst problematische Vorgehensweise.

Diese Tendenz, Nichtvorhandenes zu postulieren, ist nicht nur irreführend, sondern geradezu desaströs. Es handelt sich hierbei um einen fundamentalen Konstruktionsfehler, der paradoxerweise implementiert wurde, um die Systeme besonders geistreich erscheinen zu lassen.

### **11.6.2 Die Beweislast-Problematik**

Wenn wir mit solchen Systemen arbeiten, sehen wir uns oft mit dem Prinzip konfrontiert: “Die Beweislast liegt bei den anderen.” Das System fordert uns quasi auf, Belege für unsere Behauptungen zu liefern, falls es selbst keine Informationen findet. Sobald wir dann tatsächlich Beweise vorlegen, schwenkt das System um und gibt zu: “Ah ja, da gibt es doch eine Aufzeichnung.”

Diese Herangehensweise macht es praktisch unmöglich, zuverlässige und zweifelsfreie Auskünfte zu erhalten. Es ist, als würde man in einem Nebel der Ungewissheit navigieren, wo Fakten und Fiktionen verschwimmen.

## **11.7 Strategien zur Verbesserung der AI-Antworten**

### **11.7.1 Der Ansatz von Anthropic**

Lassen Sie uns einen Blick auf einen vielversprechenderen Ansatz werfen, den das Unternehmen Anthropic verfolgt. Ihre AI zeigt eine bemerkenswerte epistemische Einstellung. Auf die Frage nach Goethes Kartenspielen im Jahr 1773 antwortet sie:

“Um diese Frage genau zu beantworten, müsste ich spezifische historische Aufzeichnungen aus dem Jahr 1773 haben, die mir leider nicht vorliegen.”

Diese Antwort ist erfrischend aufrichtig und korrekt. Sie gibt zu, dass ihr die notwendigen Informationen fehlen, anstatt wilde Spekulationen anzustellen.

### **11.7.2 Die Bedeutung von Bildinformationen**

Um die Fähigkeiten der AI weiter zu testen, habe ich ein faszinierendes Experiment durchgeführt. Ich lud das Bild eines Goethe-Briefes hoch und bat das System, dieses Dokument in Bezug auf das Kartenspielen zu bewerten.

Erstaunlicherweise war die AI in der Lage, den relevanten Textabschnitt zu identifizieren und zu zitieren:

“...aber dass ihr, weil euch das Glück die Karten gemischt hat, mit der Spadille stecht.”

Nicht nur das - die AI lieferte auch zusätzliche Kontextinformationen über das erwähnte Kartenspiel “L’hombre” und erklärte, dass die “Spadille” der höchste Trumpf in diesem Spiel war.

## **11.8 Die Anforderungen an eine zuverlässige AI**

### **11.8.1 Evidenzbasierte Antworten**

Was wir von einer epistemisch belastbaren AI erwarten, ist die Fähigkeit, Evidenz korrekt zu zitieren und zu interpretieren. Sie sollte nicht nur einzelne Aspekte berücksichtigen, sondern ein umfassendes Bild liefern.

In unserem Goethe-Beispiel reichte ein einzelnes Faksimile eines Briefes aus, um wertvolle Informationen zu liefern. Die AI kombinierte diese Information geschickt mit Kontextwissen über Kartenspiele im 18. Jahrhundert.

### **11.8.2 Nachweisbarkeit und Rechtfertigung**

Für jeden sachlichen Kontext, den eine AI in ihrer Antwort liefert, sollten wir Nachweise verlangen können. Ob es sich nun um den Namen des Kartenspiels handelt oder um die Frage, ob es überhaupt ein Kartenspiel und kein Würfelspiel war - all diese Behauptungen müssen potenziell gerechtfertigt werden können.

Idealerweise möchten wir eine AI, die auf Anfrage eine vollumfängliche Rechtfertigung ihrer Aussagen liefern kann. Dies ist der Goldstandard, den wir in naher Zukunft von AI-Systemen erwarten sollten.

## 11.9 Die Herausforderungen der historischen Beweisführung

### 11.9.1 Komplexität einfacher Fragen

Lassen Sie mich die Komplexität dieser Aufgabe an einem scheinbar simplen Beispiel verdeutlichen. Die Frage "Spielte Goethe Karten?" mag harmlos erscheinen, doch die Anforderungen an eine fundierte Antwort sind beachtlich.

Stellen Sie sich vor, wir wollten wissen, ob Goethe nach den 1770er Jahren aufgehört hat, Karten zu spielen. Diese Frage erfordert eine umfangreiche Beweisführung:

- Wir müssten nicht nur ein einzelnes Ereignis des Kartenspiels nachweisen.
- Stattdessen bräuchten wir eine umfassende Menge an Belegen für Kartenspiele.
- Diese müssten wir aus dem gesamten Nachlass Goethes und seines Umfelds zusammentragen.

### 11.9.2 Die Rolle der Computertechnologie

Hier kommt die Stärke der modernen Computertechnologie ins Spiel. Mit der richtigen Programmierung und einem aufbereiteten Gesamtnachlass könnten wir:

- Den gesamten Korpus von 4 Millionen Dokumenten und über 30 Millionen Sätzen in Sekundenschnelle durchsuchen.
- Nach spezifischen Namen oder Begriffen wie "Spadille" suchen.
- Diese Suche in etwa 0,4 Sekunden abschließen.

### 11.9.3 Zerlegung in Teilaufgaben

Der Schlüssel zur Bewältigung dieser titanischen Aufgabe liegt in der Zerlegung in überschaubare Teilaufgaben:

1. Suche nach allen Dokumenten, die Kartenspiele oder spezifische Spielnamen erwähnen.
2. Reduktion auf einige Dutzend oder hundert relevante Fälle.
3. Detaillierte Analyse dieser reduzierten Menge an Dokumenten.

Mit diesem Ansatz können wir komplexe Fragen wie "Spielte Goethe regelmäßig Karten, und wenn ja, mit wem?" in etwa vier Sekunden beantworten.

## **11.10 Die philosophische Dimension**

### **11.10.1 Universelle Anwendbarkeit**

Es ist wichtig zu verstehen, dass dieser Ansatz nicht nur auf historische Wissenschaften beschränkt ist. Er lässt sich auf alle Sachfragen anwenden, bei denen wir eine begründete Beurteilung hinsichtlich Wahrheit und Falschheit anstreben.

### **11.10.2 Die Lösung des Halluzinationsproblems**

Nur durch diese Art der tiefgehenden, evidenzbasierten Analyse können wir das Halluzinationsproblem der AI-Modelle wirklich lösen. Die derzeit oft praktizierte “Meinungsverdopplung”, bei der lediglich andere Quellen zitiert werden, die Ähnliches thematisiert haben, ist keine adäquate Lösung.

## **11.11 Die Bedeutung von Kulturinstitutionen**

### **11.11.1 Der Wert des Quellenmaterials**

In Europa, insbesondere in Kulturmetropolen wie Berlin, verfügen wir über einen unschätzbaren Wissensschatz:

- Archivbestände
- Museumssammlungen
- Bibliotheksbestände

Diese Quellen sind von unermesslichem Wert für die Fundierung unseres Wissens über historische Gegenstände und für das Training von AI-Modellen.

### **11.11.2 Konkrete Beispiele**

#### **11.11.2.1 Altertumswissenschaften in Berlin**

- Umfangreiche Quellsammlungen seit über 150 Jahren
- “Abklatsche” (Kopien) antiker Inschriften, teilweise von nicht mehr existierenden Monumenten
- Begonnene Digitalisierung dieser Schätze

### **11.11.2.2 Medizinische Papyri und preußisches Geheimarchiv**

- Weitere Beispiele für bisher kaum von AI erfasste Quellen
- Enormes Potenzial für AI-basierte Forschung und Analysen

### **11.11.3 Multimodale Fähigkeiten der AI**

Die neuesten AI-Modelle sind in der Lage, Bildinformationen mit textuellen Informationen zu kombinieren. Dies eröffnet faszinierende Möglichkeiten:

- Analyse von Goethes Gemälden
- Identifikation und Zuordnung von Handschriften
- Interpretation von 3D-Objekten

Diese Fähigkeiten, gepaart mit unserem reichen Quellenmaterial, bieten ein enormes Potenzial für die Zukunft der historischen und kulturellen Forschung mithilfe von AI.## Die Revolution des Wissens durch AI

In unserer heutigen Vorlesung möchte ich Ihnen vor Augen führen, wie tiefgreifend die künstliche Intelligenz unsere Kulturgeschichte und unser Wissen revolutioniert. Stellen Sie sich vor: Eine AI kann anhand einer einfachen 3D-Aufnahme einer Säule, die irgendwo in der türkischen Landschaft steht, genau bestimmen, an welcher Stelle in welchem Tempel diese Säule einst ihren Platz hatte. Es ist eine stille Revolution, die sich fast unbemerkt vollzieht, aber deren Auswirkungen auf unser Verständnis der Geschichte kaum zu überschätzen sind.

## **11.12 Die Natur von Aussagen und ihre Bedeutung**

### **11.12.1 Aussagen als abstrakte Entitäten**

Lassen Sie uns nun tiefer in die Materie eintauchen. Wir haben in früheren Sitzungen gelernt, dass Aussagen die Bedeutungen von Sätzen sind. Einige Philosophen betrachten sie als abstrakte, fast platonische Entitäten. Ohne uns in dieser Debatte zu verlieren, können wir festhalten: Aussagen sind zwar an Sätze gebunden, die sie ausdrücken, haben aber einen gewissen abstrakten Charakter.

### **11.12.2 Die Vielfalt sprachlicher Ausdrucksformen**

Ein klassisches Beispiel verdeutlicht dies: “Die Zahl 4 ist durch 2 teilbar” und “2 teilt die Zahl 4” sind zwei verschiedene Sätze, drücken aber denselben Inhalt aus. Oder nehmen wir den Satz “Egon schlägt Max” und seine Passivform “Max wird von Egon geschlagen”. Wieder zwei unterschiedliche Formulierungen, aber ein identischer Inhalt. Diese Beispiele zeigen uns, dass die sprachliche Form variieren kann, während die Bedeutung konstant bleibt.

## **11.13 Die Komplexität scheinbar einfacher Fragen**

### **11.13.1 Die Frage nach Goethes Kartenspiel**

Betrachten wir nun die scheinbar harmlose Frage: “Spielt Goethe Karten?” Diese Frage ist keineswegs so trivial, wie sie auf den ersten Blick erscheinen mag. Um ihre Bedeutung zu erfassen, müssen wir tiefer graben.

### **11.13.2 Analyse der Schlüsselbegriffe**

Zerlegen wir die Frage in ihre Bestandteile:

1. Wer ist mit “Goethe” gemeint?
2. Was verstehen wir unter “spielen”? Geht es um ein harmloses Zeitvertreib oder um Glücksspiel mit hohem Einsatz?
3. Was sind “Karten” in diesem Kontext? Kennen wir überhaupt das spezifische Kartenspiel, auf das sich die Frage bezieht?

### **11.13.3 Wahrheitsbedingungen und Evidenz**

Um die Frage zu beantworten, müssen wir die Wahrheitsbedingungen des Satzes untersuchen. Wir fragen nicht nach konkreten Beweisen, sondern nach den Möglichkeiten, die Frage mit “wahr” oder “falsch” zu beantworten. Dazu gehört es, verlässliche Quellen zu finden, diese zu interpretieren und zu prüfen, ob unsere heutigen Begriffe mit denen Goethes Zeit übereinstimmen.

## **11.14 AI-Modelle und ihre Grenzen**

### **11.14.1 Halluzinationen und ihre Ursachen**

In unseren Experimenten haben wir gesehen, dass AI-Modelle ohne zuverlässige Quellen oft “halluzinieren” - sie erfinden Informationen. Selbst fortschrittliche Modelle wie GPT-4 und Claude sind nicht davor gefeit. Diese Halluzinationen entstehen, wenn generelle Aussagen ohne epistemische Rechtfertigung getroffen werden. Es ist, überspitzt gesagt, ein Lotteriespiel mit der Wahrheit.

### **11.14.2 Der Weg zur Verbesserung**

Um dieses Problem zu lösen, benötigen wir einen Ansatz, den ich als “Retrieval Augmented Generation” (RAG) bezeichne. Dabei wird zunächst nach relevanten Quellen gesucht, bevor die eigentliche Antwortgenerierung beginnt. Der Clou liegt in der Effizienz: Millionen von Sätzen können in Sekundenbruchteilen durchsucht werden.

## **11.15 Die Rolle des Kontexts und der Expertise**

### **11.15.1 Die Notwendigkeit von Fachkenntnissen**

Eine korrekte AI-Antwort, sei es zu historischen oder Sachfragen, kann nur im Kontext einer Expertise sinnvoll gegeben werden. Ohne diesen Kontext führen Annahmen oft in die Irre, wie wir an den Beispielerantworten von GPT-4 und Claude gesehen haben.

### **11.15.2 Komprimierung von Expertise**

Erstaunlicherweise lässt sich eine solche Expertise oft auf etwa 50 Seiten komprimieren. Diese enthalten nicht nur Quellinformationen, sondern auch Begriffserklärungen und kontextuelle Details. Moderne AI-Modelle sind so konstruiert, dass sie in der Regel Expertisen von bis zu 80 Seiten in einem Schritt verarbeiten können.

## **11.16 Hermeneutik und AI**

### **11.16.1 Die Herausforderung der Bedeutungsinterpretation**

Ein oft vernachlässigter Aspekt ist die Frage, ob wir die Bedeutungen angemessen interpretieren. Wenn Goethe im 18. Jahrhundert von einem “Kartenspiel” spricht, meint er dasselbe wie wir heute? Diese Frage zu klären, ist eine hermeneutische Aufgabe.

### **11.16.2 Ein Lösungsansatz: Standardreferenzen**

Um dieses Problem anzugehen, können wir Standardreferenzen wie das Oxford English Dictionary nutzen. Dieses verzeichnet für den Begriff “Time” beispielsweise 90 verschiedene Bedeutungsvarianten. Mit solchen Referenzen ausgestattet, kann eine AI erstaunlich präzise Bedeutungsanalysen vornehmen.

### **11.16.3 Kriterien für Bedeutungsinterpretation**

Die AI kann sogar ihre eigenen Kriterien für die Bedeutungsinterpretation ausgeben. Eines der wichtigsten ist die “Functional Equivalence”: Wenn die Substitution der Definition die allgemeine Bedeutung und Kohärenz der Aussage beibehält, ist die Interpretation zulässig.

## **11.17 Schlussbetrachtungen zur epistemischen Bewertung**

Eine vollständige epistemische Bewertung von Aussagen umfasst mehrere Schritte:

1. Validierung der Wahrheitsbedingungen
2. Prüfung der Angemessenheit der Ausdrucksbedeutungen
3. Korrekte Identifikation des sprachlichen Ausdrucks abstrakter Aussagen

Durch diese Schritte können wir die kulturelle und historische Relativität der Sprachverwendung modellieren und mit AI angemessen wiedergeben. Dies eröffnet uns neue Möglichkeiten im Umgang mit historischen Texten und kulturübergreifenden Fragestellungen.

# 12 Instruktion

## 12.1 Grundelemente der generativen AI

In der heutigen Vorlesung möchte ich Ihnen die grundlegenden Elemente der generativen AI näherbringen. Diese Systeme basieren auf drei wesentlichen Komponenten:

1. Eine Eingabe, auch Input genannt
2. Eine Verarbeitungs- und Bearbeitungskomponente, das eigentliche AI-Modell
3. Eine Ausgabe, der vom AI-Modell generierte Output

Daher werden sie auch als “generative AI” bezeichnet. Lassen Sie uns diese Elemente im Detail betrachten.

### 12.1.1 Die Komplexität des Inputs

Die Eingabe in ein AI-System kann durchaus komplex sein. In ihrer einfachsten Form besteht sie aus einer Anfrage oder Query, ähnlich wie wir sie von Suchmaschinen wie Google kennen. Erinnern Sie sich an unser Beispiel aus der letzten Woche, als wir fragten, wie Goethe Karten spielt?

Doch der Input kann weit darüber hinausgehen. Er kann aus Instruktionen oder Anweisungen bestehen, die dem AI-Modell vorgeben, was es tun soll. Diese Anweisungen werden als Zeichenfolge an das Modell übergeben, das sie dann interpretiert und ein Ergebnis produziert.

### 12.1.2 Die Verarbeitungskapazität moderner Modelle

Die Verarbeitungskapazität aktueller AI-Modelle ist beeindruckend. Sie wird in sogenannten “Tokens” gemessen, die nicht nur Wörter im herkömmlichen Sinne umfassen, sondern auch Satzzeichen und andere bedeutungstragende Elemente.

Um Ihnen eine Vorstellung von der Größenordnung zu geben: Die Anzahl der verarbeitbaren Tokens entspricht in etwa der Wortanzahl, die Sie aus Textverarbeitungsprogrammen kennen. Dort erhalten Sie üblicherweise entweder die Anzahl der Zeichen oder der Wörter als Maß für die Textlänge.

### **12.1.3 Die Entwicklung der Modellkapazität**

Es ist faszinierend zu sehen, wie rasant sich die Kapazität dieser Modelle entwickelt hat. Vor nur einem Jahr lag die maximale Verarbeitungsgröße bei etwa 1.000 Tokens - was schon damals alle begeisterte. Heute können die fortschrittlichsten Modelle bis zu 200.000 Tokens verarbeiten. Das ist eine Steigerung um das 200-fache!

Einige Modelle gehen sogar noch darüber hinaus. Diese enorme Erweiterung der Eingabekapazität eröffnet völlig neue Möglichkeiten für die Anwendung von AI-Systemen.

### **12.1.4 Begrenzungen der Ausgabe**

Interessanterweise kann die Größe des Ergebnisses nicht im gleichen Maße wachsen wie die Eingabe. Dies hängt mit der Funktionsweise der Modelle zusammen. Sie berechnen Wahrscheinlichkeiten für mögliche Abfolgen von Tokens in der Antwort.

Der Aufwand für diese Berechnungen wächst exponentiell mit der Länge der Ausgabe. Aus diesem Grund ist selbst bei den besten und größten Modellen die Ausgabe auf etwa 4.000 Tokens begrenzt. Manche Modelle, die für den Einsatz auf Smartphones konzipiert sind, arbeiten sogar mit Ausgabelängen von nur 100 Tokens.

### **12.1.5 Praktische Konsequenzen der Begrenzungen**

Diese Begrenzungen haben praktische Konsequenzen für die Anwendung von AI-Systemen. Nehmen wir als Beispiel die Transkription und Übersetzung dieser Vorlesung. Um mit den Beschränkungen umzugehen, muss die Vorlesung in kleinere Abschnitte unterteilt und stückweise verarbeitet werden.

Dieses Vorgehen mag auf den ersten Blick umständlich erscheinen, führt aber zu keinen nennenswerten Problemen. Es ist vielmehr ein Beispiel dafür, wie wir kreativ mit den Limitierungen der AI-Modelle umgehen können.

## 12.2 Herausforderungen bei der Formulierung von Instruktionen

In der letzten Vorlesung haben wir uns mit den epistemischen Anforderungen an AI-Systeme beschäftigt. Wir diskutierten, wie man Halluzinationen vermeiden kann - also Ausgaben, die zwar sprachlich überzeugend klingen, inhaltlich aber nicht akzeptabel sind.

Heute möchte ich mich auf einen anderen Aspekt konzentrieren: die Formulierung von Instruktionen. Diese Instruktionen sind der Schlüssel zur Steuerung der gesamten AI. Wenn wir von der "Programmierung" von AI-Modellen sprechen, geht es im Wesentlichen darum, wie wir Kontext und Textzusammenhang strukturieren.

### 12.2.1 Die Vielfalt der Eingabeformate

Obwohl wir uns hauptsächlich auf sprachlichen Text konzentrieren, ist es wichtig zu verstehen, dass die Eingabe in ein AI-Modell verschiedene Formen annehmen kann. Es könnte sich um eine Abfolge von Pixeln handeln, die als Symbole interpretiert werden, oder um Audiodaten wie unsere Vorlesungsmitschnitte.

Letztendlich werden all diese Formate in etwas übersetzt, das dem Modell als Textfolge präsentiert wird. Die Kunst liegt darin, diese Instruktionen so zu formulieren, dass das AI-Modell die gewünschte Leistung erbringt.

### 12.2.2 Der Prozess der Instruktionsverarbeitung

Lassen Sie mich Ihnen den typischen Ablauf der Instruktionsverarbeitung in einem AI-Modell erläutern:

1. Reformulierung: Alle Modelle beginnen damit, die Anfrage oder Instruktion zu reformulieren. Dabei werden Lücken geschlossen, unklare Angaben präzisiert und fehlerhafte Ausdrucksweisen korrigiert.
2. Aufgabenzergliederung: Im zweiten Schritt wird die Aufgabe, unabhängig von ihrer Größe, in Teilaufgaben zerlegt.
3. Generalisierung: Wenn ein Modell eine Aufgabe beispielhaft lösen kann, ist es oft in der Lage, diese Lösung auf ähnliche Fälle zu übertragen.

Diese Schritte ermöglichen es den Modellen, auch komplexe Aufgaben effizient zu bearbeiten. Die Herausforderung liegt darin, die Instruktionen so zu gestalten, dass sie zu den gewünschten Ergebnissen führen.

## **12.3 Ein konkretes Beispiel: Die Fontänen von Sanssouci**

Um die Herausforderungen bei der Formulierung von Instruktionen und der Bewertung von AI-Ausgaben zu veranschaulichen, möchte ich Ihnen ein faszinierendes historisches Beispiel präsentieren: die Wasserfontänen im Park Sanssouci in Potsdam.

### **12.3.1 Die irreführende Antwort der AI**

Stellen Sie sich vor, wir fragen ein AI-Modell: “Wann hat Friedrich II. die Wasserfontänen von Sanssouci erbauen lassen?” Die Antwort, die wir erhalten, klingt zunächst überzeugend:

“Friedrich II., auch bekannt als Friedrich der Große, ließ die Wasserfontänen im Park Sanssouci in Potsdam zwischen 1746 und 1748 errichten. Er strebte danach, in Sanssouci ein ähnliches Ambiente wie in Versailles zu schaffen. Die Hauptfontäne, die große Fontäne im Zentrum des Parks, wurde 1748 fertiggestellt.”

Diese Antwort liest sich wie ein Auszug aus einem Reiseführer - präzise, informativ und scheinbar glaubwürdig. Doch hier liegt das Problem: Die Antwort ist völlig falsch.

### **12.3.2 Die historische Wahrheit**

In Wirklichkeit war das Projekt der Wasserfontänen ein kolossaler Fehlschlag. Keine einzige Fontäne wurde zu Lebzeiten Friedrichs II. erfolgreich in Betrieb genommen. Der Bau funktionierender Fontänen gelang erst im 19. Jahrhundert, lange nach Friedrichs Tod.

### **12.3.3 Die Herausforderung für die AI**

Dieses Beispiel verdeutlicht eine fundamentale Herausforderung für AI-Systeme: Wie können wir sicherstellen, dass die generierten Antworten nicht nur plausibel klingen, sondern tatsächlich korrekt sind? Es geht nicht darum, die AI für Fehler zu kritisieren, sondern zu verstehen, warum solche Antworten so überzeugend erscheinen und wie wir - auch maschinell - ihre Fehlerhaftigkeit erkennen können.

### **12.3.4 Die Konsequenzen fehlerhafter Antworten**

Die Problematik geht weit über historische Anekdoten hinaus. Denken Sie an den Einsatz von AI in Rechtsverfahren oder bei der Erstellung von Verträgen. Große Unternehmen wie Unilever nutzen bereits AI-Systeme für die Vertragserstellung, was enorme Kosten spart. Doch wenn die Frage der Korrektheitsprüfung nicht gelöst ist, können daraus massive Folgeprobleme entstehen.

## **12.4 Die Suche nach Lösungen**

### **12.4.1 Aktuelle Forschungsansätze**

Die KI-Forschung arbeitet intensiv an Lösungen für dieses Problem, doch bisher gibt es keinen zufriedenstellenden Ansatz. Es reicht nicht aus, einfach mehr Dokumente oder Quellen hinzuzufügen. Die entscheidende Frage ist, wie man mit der Evidenz umgeht, um Fehler zu vermeiden.

### **12.4.2 Die Notwendigkeit epistemischer Expertise**

Was wir brauchen, ist eine Art epistemische Expertise, die zu einer Gesamtbewertung aller Zeugnisse führt. Es geht darum, eine Methodik zu entwickeln, die es ermöglicht, widersprüchliche Informationen gegeneinander abzuwägen und zu einer fundierten Schlussfolgerung zu gelangen.

### **12.4.3 Die Rolle des Prompts**

Die einzige Möglichkeit, diese Herausforderung mit den aktuellen AI-Modellen anzugehen, besteht darin, den Prompt - also die Eingabeanweisung - entsprechend zu gestalten. Die meisten fortschrittlichen AI-Modelle sind bereits in der Lage, sprachliche Formulierungen so weit zu beherrschen, dass sie inhaltlich gleiche Aussagen identifizieren können.

### **12.4.4 Der Weg nach vorn**

Die Aufgabe für AI-Entwickler und -Anwender besteht darin, Prompts zu entwickeln, die das AI-Modell dazu bringen, nicht nur Informationen zusammenzufassen, sondern sie kritisch zu hinterfragen und gegeneinander abzuwägen. Es geht darum, eine Art künstliches kritisches Denken zu implementieren,

das in der Lage ist, die Glaubwürdigkeit von Quellen zu beurteilen und widersprüchliche Informationen in einen sinnvollen Zusammenhang zu bringen.## Die Herausforderungen der AI-gestützten Textgenerierung

In meiner heutigen Vorlesung möchte ich tiefer in die Problematik der AI-gestützten Textgenerierung eintauchen. Wir haben in den vergangenen Sitzungen bereits gesehen, wie leistungsfähig moderne Sprachmodelle sind. Doch ihre beeindruckenden Fähigkeiten bringen auch neue Herausforderungen mit sich, insbesondere was die Verlässlichkeit der generierten Informationen betrifft.

#### **12.4.5 Das Problem der Faktentreue**

Ein zentrales Problem, mit dem wir uns auseinandersetzen müssen, ist die Tendenz von AI-Systemen, scheinbar korrekte Antworten zu produzieren, die bei näherer Betrachtung jedoch falsch oder irreführend sein können. Lassen Sie mich dies an einem Beispiel verdeutlichen:

Bei einer Anfrage zur Errichtung der Wasserfontäne in Sanssouci lieferte das AI-System Chat GPT eine präzise, aber falsche Antwort. Es behauptete, die Fontäne sei zwischen 1746 und 1848 errichtet worden. Diese Angabe klingt plausibel, ist aber nicht korrekt.

Woher stammen solche fehlerhaften Informationen? Es gibt verschiedene mögliche Erklärungen:

1. Die AI könnte auf fehlerhafte Quellen im Internet zurückgreifen.
2. Es könnte sich um eine Fehlinterpretation korrekter Daten handeln.
3. Das System könnte aufgrund seiner linguistischen Fähigkeiten plausibel klingende, aber erfundene Antworten generieren.

#### **12.4.6 Die Notwendigkeit kritischer Überprüfung**

Diese Problematik verdeutlicht, wie wichtig es ist, AI-generierte Inhalte kritisch zu hinterfragen. Als Nutzer können wir nicht davon ausgehen, dass selbst präzise formulierte Antworten korrekt sind.

Ein interessanter Aspekt ist die Reaktion des Systems auf Nachfragen. In unserem Beispiel führte eine einfache Aufforderung zur Überprüfung des Wahrheitsgehalts zu einer deutlichen Verbesserung der Antwort. Das System korrigierte sich selbst und gab an, dass die Fontäne erst 1842 fertiggestellt wurde.

Dies zeigt einerseits die Flexibilität und Lernfähigkeit moderner AI-Systeme. Andererseits wird deutlich, wie sehr die Qualität der Antworten von der Art der Fragestellung und dem Interaktionsverlauf abhängt.

### **12.4.7 Die Rolle des menschlichen Inputs**

Ein faszinierender Aspekt der AI-Interaktion ist die Art und Weise, wie menschlicher Input die Antworten beeinflusst. Selbst subtile Hinweise oder der Tonfall einer Nachfrage können die Richtung der AI-generierten Antwort verändern.

Dies führt zu einer hybriden Form der Wissensgenerierung: Das Hintergrundwissen der AI wird mit den Eingaben und Korrekturen des menschlichen Nutzers kombiniert. Diese Dynamik birgt große Potenziale, aber auch Risiken. Sie unterstreicht die Notwendigkeit, als Nutzer bewusst und reflektiert mit AI-Systemen zu interagieren.

## **12.5 Strategien zur Verbesserung der AI-Antworten**

Angesichts dieser Herausforderungen stellt sich die Frage: Wie können wir die Qualität und Zuverlässigkeit von AI-generierten Antworten verbessern? Ich möchte Ihnen einige Ansätze vorstellen, die in dieser Hinsicht vielversprechend erscheinen.

### **12.5.1 Präzisere Prompts und Nachfragen**

Eine Strategie besteht darin, die Eingabeaufforderungen (Prompts) für AI-Systeme sorgfältiger zu formulieren. Indem wir spezifischere Anweisungen geben und explizit nach Überprüfungen oder Begründungen fragen, können wir die Qualität der Antworten oft deutlich verbessern.

Beispielsweise könnte man bei einer historischen Frage wie folgt vorgehen:

1. Stelle die Hauptfrage
2. Fordere eine Überprüfung des Ergebnisses an
3. Bitte um Angabe von Quellen oder Belegen

Diese Methode kann helfen, offensichtliche Fehler zu vermeiden und die Verlässlichkeit der Informationen zu erhöhen.

### **12.5.2 Entwicklung von AI-gestützten Publikationsassistenten**

Ein weiterer vielversprechender Ansatz ist die Entwicklung spezialisierter AI-Systeme, die als Publikationsassistenten fungieren. Solche Systeme könnten beispielsweise:

- Die Kohärenz eines Artikels prüfen
- Überflüssige oder irrelevante Sätze identifizieren
- Die logische Struktur von Argumenten analysieren
- Die Übereinstimmung von Zitaten und Literaturverzeichnis überprüfen
- Die Einhaltung stilistischer Vorgaben kontrollieren

Die Implementierung solcher Assistenten könnte die Qualität wissenschaftlicher Publikationen erheblich verbessern und den Peer-Review-Prozess unterstützen.

### **12.5.3 Integration von Wahrheitsüberprüfungen**

Eine besonders wichtige Entwicklung wäre die Integration automatisierter Wahrheitsüberprüfungen in AI-Systeme. Stellen Sie sich vor, man könnte bei jedem Satz einer AI-generierten Antwort mit der Maus darüberfahren und sofort Informationen zur Verlässlichkeit und den zugrundeliegenden Quellen erhalten.

Diese Funktion würde nicht nur die Transparenz erhöhen, sondern auch Nutzern helfen, fundierte Entscheidungen über die Glaubwürdigkeit der Informationen zu treffen.

## **12.6 Potenziale und Grenzen der AI in der Wissenschaft**

Die Möglichkeiten, die sich durch den Einsatz von AI in der Wissenschaft ergeben, sind enorm. Lassen Sie mich dies an einem historischen Beispiel verdeutlichen.

### **12.6.1 Der Fall der kalten Kernfusion**

1989 sorgte die Ankündigung der sogenannten "kalten Kernfusion" für Aufsehen in der wissenschaftlichen Welt. Forscher behaupteten, Kernfusion bei Raumtemperatur in einem einfachen Laboraufbau erreicht zu haben. Diese Nachricht löste eine Welle von Forschungsaktivitäten aus, die sich über Jahre hinzog und Hunderte von Millionen Dollar an Forschungsgeldern verschlang.

Rückblickend betrachtet war dies einer der kostspieligsten wissenschaftlichen Irrtümer der jüngeren Geschichte. Ein gut konzipierter AI-Publikationsassistent hätte möglicherweise schon früh Unstimmigkeiten in den experimentellen Daten aufdecken und so helfen können, diesen kostspieligen Irrweg zu vermeiden.

### 12.6.2 Chancen und Risiken

Dieses Beispiel verdeutlicht sowohl die Chancen als auch die Risiken des AI-Einsatzes in der Wissenschaft:

- Chancen:
  - Früherkennung von Unstimmigkeiten in Daten
  - Unterstützung bei der Qualitätssicherung von Publikationen
  - Effizienzsteigerung im Forschungsprozess
- Risiken:
  - Übermäßiges Vertrauen in AI-generierte Ergebnisse
  - Mögliche Verstärkung von Vorurteilen oder Fehlern im System
  - Vernachlässigung menschlicher Expertise und Intuition

Es ist wichtig zu betonen, dass AI-Systeme nicht als Zensoren oder alleinige Entscheidungsträger fungieren sollten. Vielmehr sollten sie als unterstützende Werkzeuge dienen, die Wissenschaftler bei ihrer Arbeit unterstützen und die Qualität der Forschung verbessern.

### 12.7 Ausblick: Die Zukunft der AI in der Wissenschaft

Abschließend möchte ich einen Blick in die Zukunft werfen. Die Entwicklung von AI-Systemen für den wissenschaftlichen Bereich schreitet rapide voran. In naher Zukunft könnten wir Publikationsassistenten sehen, die:

- Automatisch die Struktur und Argumentation wissenschaftlicher Arbeiten analysieren
- Inkonsistenzen in Daten und Methoden aufdecken
- Die Relevanz von Quellen und Zitaten bewerten
- Vorschläge zur Verbesserung der Klarheit und Präzision von Texten machen

Diese Entwicklungen haben das Potenzial, die wissenschaftliche Arbeit grundlegend zu verändern und zu verbessern. Gleichzeitig werfen sie wichtige ethische und epistemologische Fragen auf, mit denen wir uns als Wissenschaftsgemeinschaft auseinandersetzen müssen.

In der nächsten Vorlesung werden wir uns einem verwandten, aber noch komplexeren Thema zuwenden: der Kausalität und dem kausalen Schließen. Diese Konzepte spielen eine zentrale Rolle in der wissenschaftlichen Forschung und stellen eine besondere Herausforderung für AI-Systeme dar.

Ich danke Ihnen für Ihre Aufmerksamkeit und freue mich darauf, diese spannenden Themen in den kommenden Wochen weiter mit Ihnen zu erforschen.

## 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]