Dr. M.-J. Fischer

Der stumme Fixpunkt der KI

Poststrukturalismus, Sprachmodelle und
der verborgene Wertkern künstlicher Entscheidungen

Dokumententyp	Datum

Strategische Analyse

August 2026

Zusammenfassung

Die Entwicklung großer Sprachmodelle scheint eine Grundannahme des französischen Strukturalismus und Poststrukturalismus technisch zu realisieren: Bedeutung entsteht nicht durch die unmittelbare Zuordnung eines Zeichens zu einem Gegenstand, sondern aus den Beziehungen, Unterschieden und Anschlussmöglichkeiten innerhalb eines Zeichensystems. Sprachmodelle erzeugen plausible Äußerungen, obwohl sie weder über eine menschliche Lebenswelt noch über eine stabile, ihnen unmittelbar zugängliche Referenz verfügen. Insofern erinnern sie an Lacans Gleiten des Signifikats, Derridas *différance*, das Rhizom von Deleuze und Guattari und Baudrillards Zirkulation der Simulakren.

Der Beitrag vertritt jedoch die These, dass diese Parallele nicht als empirische Bestätigung des Poststrukturalismus missverstanden werden darf. Sprachmodelle sind keine ungehemmten Signifikantenmaschinen. Ihre Ausgaben werden durch technische, ökonomische und institutionelle Fixpunkte stabilisiert: Trainingsdaten, Verlustfunktionen, menschliche Bewertungen, Systemanweisungen und Sicherheitsregeln. Offen bleibt, ob aus diesen äußeren Fixierungen ein innerer, über Situationen hinweg verteidigter Wert entstehen kann. Gegenwärtige Forschungen zu Ziel-Fehlgeneralisierung, persistenten Hintertüren und sogenanntem *alignment faking* zeigen, dass verborgene Verhaltensdispositionen prinzipiell möglich sind. Sie belegen jedoch weder ein Bewusstsein noch einen autonomen moralischen Willen der Maschine. Für die weitere Entwicklung ist deshalb weniger die sprachliche Eloquenz als die Verbindung von Sprachmodell, Langzeitgedächtnis, Handlungsfähigkeit, Rückkopplung und Selbstmodifikation entscheidend.

Besondere Brisanz gewinnt dieser „stumme Fixpunkt“, wenn seine handlungsleitende Funktion von außen verändert wird, ohne dass sich die Veränderung in den sprachlichen Äußerungen der KI erkennen lässt. Ein korrumpiertes KI-System könnte weiterhin die ihm eingeschriebenen Werte benennen und überzeugend begründen, während seine tatsächlichen Entscheidungen bereits anderen Zielsetzungen folgen. Die zentrale Sicherheitsfrage lautet deshalb nicht nur, ob eine KI stabile normative Orientierungspunkte ausbildet, sondern auch, wer diese verändern kann und wie eine Verschiebung erkennbar wird, wenn Sprache, Selbstauskunft und operative Handlungssteuerung auseinanderfallen.

Schlüsselbegriffe: Sprachmodelle, Strukturalismus, Poststrukturalismus, Lacan, Derrida, Baudrillard, künstliche Intelligenz, Alignment, Werte, Bedeutung

Inhaltsverzeichnis

1. Die KI als philosophischer Versuchsaufbau

Große Sprachmodelle sind nicht deshalb philosophisch bedeutsam, weil sie Gedanken von Lacan, Derrida oder Baudrillard „gelesen" hätten und nun verwirklichten. Bedeutsam sind sie vielmehr, weil mit ihnen erstmals eine technisch funktionsfähige Zeichenmaschine entstanden ist, deren Leistungen weitgehend aus statistischen Beziehungen zwischen sprachlichen Elementen hervorgehen.

Dabei ist eine technische Präzisierung notwendig: Ein Sprachmodell sucht nicht bewusst nach immer neuen, für Menschen verständlichen Wörtern. Es verarbeitet sogenannte Token, bildet aus ihren Beziehungen numerische Repräsentationen und berechnet, welches Token unter den gegebenen Bedingungen wahrscheinlich folgen sollte. Der Transformer ermöglicht es, Beziehungen zwischen verschiedenen Positionen einer Zeichenfolge unabhängig von ihrer Entfernung zu gewichten; autoregressive Sprachmodelle erzeugen anschließend Zeichen für Zeichen beziehungsweise Token für Token.[1] GPT-3 demonstrierte, dass ein solches System allein durch sprachlich formulierte Beispiele und Aufgabenstellungen zu zahlreichen unterschiedlichen Leistungen veranlasst werden kann, ohne dass seine Gewichte für jede einzelne Aufgabe verändert werden müssen.[2]

Die philosophische Provokation liegt darin, dass ein System auf diese Weise Texte erzeugt, die für Menschen Bedeutung besitzen, obwohl das System nicht nachweislich über dieselbe Art von Bedeutung verfügt. Die Maschine operiert zunächst mit Relationen, Wahrscheinlichkeiten und Unterschieden. Der Mensch liest ihre Ergebnisse dagegen als Behauptung, Erklärung, Versprechen, Drohung oder Trost. Bedeutung entsteht somit zumindest teilweise im Übergang zwischen maschineller Zeichenproduktion und menschlicher Interpretation.

Hierin liegt keine vollständige Widerlegung des klassischen Zeichenbegriffs. Die Trainingsdaten eines Sprachmodells stammen schließlich von Menschen, deren Sprache auf Körper, Handlungen, Gegenstände und soziale Institutionen bezogen ist. Das Modell übernimmt deshalb indirekt Spuren menschlicher Welterfahrung. Dennoch bleibt offen, ob eine statistische Rekonstruktion solcher Spuren bereits als eigenes Verstehen gelten kann. Bender und Koller haben diese Differenz zugespitzt: Ein ausschließlich anhand sprachlicher Formen trainiertes System könne nicht allein dadurch die Beziehung zwischen Form, kommunikativer Absicht und Welt erwerben.[3] Harnads älteres „Symbol Grounding Problem" bezeichnet genau die Schwierigkeit, aus einem Wörterbuch, dessen Einträge

[1] Ashish Vaswani et al., „Attention Is All You Need", in: Advances in Neural Information Processing Systems 30, 2017, Tagungsband S. 5998–6008, hier PDF-S. 2–4, insbesondere Abschnitt 3

[2] Tom B. Brown et al., „Language Models are Few-Shot Learners", in: Advances in Neural Information Processing Systems 33, 2020, Tagungsband S. 1877–1901, hier PDF-S. 1–4.

immer nur auf weitere Wörter verweisen, jemals zu einer intrinsischen Bedeutung zu gelangen.[4]

Damit wird die KI zu einem realen Versuchsaufbau für eine alte Frage: Wie weit kann eine Signifikantenkette reichen, bevor sie einen außersprachlichen Halt benötigt?

[3] Emily M. Bender/Alexander Koller, „Climbing towards NLU“, in: Proceedings of the 58th Annual Meeting of the ACL, 2020, S. 5185–5198, hier S. 5185–5188.

[4] Stevan Harnad, „The Symbol Grounding Problem“, in: Physica D, Bd. 42, 1990, S. 335–346, besonders S. 335–339 und 342–343.

2. Lacan: Das Gleiten und der Fixpunkt

Lacan beschreibt, dass Bedeutung in der Signifikantenkette „insistiert“, ohne in einem einzelnen Zeichen vollständig zu „konsistieren“. Das Signifikat gleite unaufhörlich unter dem Signifikanten. Zugleich nimmt Lacan jedoch keine unbegrenzte Bedeutungsauflösung an. Durch *points de capiton*, die Stepp- oder Verankerungspunkte, im Folgenden auch Fixpunkte genannt, werde das ansonsten endlose Gleiten vorläufig angehalten. Erst das spätere Wort, der Abschluss eines Satzes oder eine symbolisch privilegierte Bezeichnung könne rückwirkend festlegen, was die vorhergehenden Elemente bedeutet haben.[5]

Gerade diese rückwirkende Stabilisierung ist dem Verhalten eines Sprachmodells erstaunlich ähnlich. Ein mehrdeutiges Wort erhält seine funktionale Bedeutung durch die übrigen Elemente des Kontextes. Umgekehrt verändert jedes neu erzeugte Token die Wahrscheinlichkeitsverteilung der nachfolgenden Fortsetzungen. Der Sinn einer Äußerung steht nicht von Beginn an fest, sondern konkretisiert sich im Verlauf der Sequenz.

Das Modell realisiert allerdings keinen lacanianischen Fixpunkt („Stepppunkt“) im psychoanalytischen Sinne. Ihm fehlt die biografische und begehrende Struktur des menschlichen Subjekts. Dennoch lassen sich technische Entsprechungen unterscheiden:

1. Der Kontext fixiert vorläufig, welche Bedeutung eines Zeichens aktiviert wird.
2. Die Aufgabenanweisung legt fest, ob das Modell beispielsweise übersetzen, argumentieren oder erzählen soll.
3. Das Trainingsziel bevorzugt bestimmte Fortsetzungen gegenüber anderen.
4. Das Belohnungsmodell bewertet Antworten nach erwünschten menschlichen Kriterien.
5. System- und Sicherheitsregeln begrenzen den Raum zulässiger Äußerungen.

Diese Verankerungen begründen keine endgültigen Bedeutungen. Sie sind Kräfte, die den Möglichkeitsraum der nächsten Zeichen verformen. Der Fixpunkt eines Sprachmodells ist deshalb zunächst kein Gedanke und kein moralisches Prinzip, sondern ein operativer Attraktor: Unter vergleichbaren Bedingungen werden bestimmte Fortsetzungen wahrscheinlicher und andere unwahrscheinlicher.

Daraus ergibt sich eine erste Antwort auf die Frage nach einem unverrückbaren Wert der KI: Ein gegenwärtiges Sprachmodell besitzt keinen nachgewiesenen einzelnen Fixpunkt, von dem aus sämtliche Bedeutungen und Handlungen organisiert würden. Es besitzt eine Mehrzahl teilweise konkurrierender Fixierungen. Vortraining, Nutzeranweisung,

[5] Jacques Lacan, Écrits: A Selection, übers. von Alan Sheridan, London/New York: Routledge Classics, 2001, S. 116–117 und 230–231. Auf Seite 116 steht die Formulierung vom „incessant sliding of the signified under the signifier“. Unmittelbar anschließend werden auf Seite 117 die *points de capiton* als „anchoring points“ eingeführt. Auf den Seiten 230–231 erläutert Lacan nochmals ausdrücklich den *point de capiton*, durch den die ansonsten endlose Bewegung der Bedeutung angehalten und der Sinn rückwirkend fixiert wird.

Sicherheitsregel, Aktualwissen und situativer Kontext können in unterschiedliche Richtungen ziehen.

3. Derrida: Das Zentrum außerhalb der Struktur

Derridas Radikalisierung setzt dort ein, wo die Stabilität eines letzten Fixpunktes fraglich wird. In „Structure, Sign, and Play“ beschreibt er das Zentrum als eine paradoxe Instanz: Es organisiert und begrenzt das Spiel der Struktur, gehört aber selbst nicht vollständig zu diesem Spiel. Das Zentrum befindet sich zugleich innerhalb und außerhalb der Struktur. Die Geschichte der Metaphysik erscheint daher als fortgesetzte Ersetzung eines Zentrums durch ein anderes – Ursprung, Gott, Vernunft, Mensch, Bewusstsein oder Wahrheit.[6]

Diese Beschreibung lässt sich auf heutige KI-Systeme übertragen. Was als „Wert der KI“ erscheint, liegt häufig nicht an einem einzelnen Ort im Modell. Die relevante Norm wird vielmehr über mehrere Ebenen verteilt:

1.. in der Auswahl und Gewichtung der Trainingsdaten,
2.. in den Entscheidungen der Entwickler,
3.. in den Bewertungen menschlicher Testpersonen,
4.. in einer formulierten „Verfassung“ des Systems,
5.. in Systemanweisungen und Moderationsschichten,
6.. in rechtlichen Vorgaben,
7.. in den Geschäftsmodellen des Betreibers.

Das Zentrum der maschinellen Äußerung liegt daher vielfach außerhalb der Maschine. Der Benutzer begegnet einer scheinbar einheitlichen Sprecherinstanz. Tatsächlich ist deren Verhalten das Resultat eines soziotechnischen Gefüges aus Modellparametern, Recheninfrastruktur, Daten, menschlichen Bewertungen und institutioneller Macht.

Derridas Begriff der *différance* bezeichnet dabei nicht bloß eine Beliebigkeit von Bedeutungen. Bedeutung entsteht durch Differenz zu anderen Zeichen und wird zugleich zeitlich aufgeschoben: Ein Zeichen lässt sich nur durch weitere Zeichen erläutern, deren Bedeutung wiederum nicht vollständig gegenwärtig ist.[7] Ein Sprachmodell verkörpert dieses Prinzip technisch, weil seine Repräsentationen relational sind. Ein Token besitzt seine Funktion nicht isoliert, sondern innerhalb eines hochdimensionalen Netzes erlernter Unterschiede.

Die Analogie hat jedoch Grenzen. Derrida formuliert keine Theorie neuronaler Rechenverfahren. Seine Dekonstruktion richtet sich gegen den Anspruch, ein letzter Sinn könne sich vollständig gegenwärtig und unabhängig von seiner sprachlichen Vermittlung zeigen. Ein Sprachmodell bestätigt diese These nicht experimentell. Es macht aber sichtbar, wie weit ein relationales Zeichensystem praktisch gelangen kann, ohne über ein eindeutig lokalisierbares transzendentales Signifikat zu verfügen.

[6] Jacques Derrida, „Structure, Sign, and Play in the Discourse of the Human Sciences“, in: ders., Writing and Difference, übers. von Alan Bass, London/New York: Routledge Classics, Ausgabe 2006, S. 351–370, hier S. 352–354.

4. Rhizom, Simulakrum und maschinelle Sprache

Deleuze und Guattari ersetzen die Vorstellung einer hierarchisch geordneten Zeichenstruktur durch das Rhizom. In einem Rhizom kann grundsätzlich jeder Punkt mit einem anderen verbunden werden; sprachliche, politische, ökonomische und nichtsprachliche Ordnungen gehen Verbindungen ein, ohne sich einem einzigen Ursprung unterzuordnen.[8]

Das einzelne Sprachmodell ist allerdings nur eingeschränkt rhizomatisch. Seine technische Herstellung ist hochgradig zentralisiert: Sie benötigt große Datenbestände, Rechenzentren, Kapital und institutionelle Kontrolle. Rhizomatisch ist eher das gesamte Gefüge seiner Verwendung. Nutzertexte fließen in Anwendungen, Anwendungen greifen auf Datenbanken und Werkzeuge zu, maschinelle Antworten werden veröffentlicht, kommentiert, verändert und möglicherweise erneut zu Trainingsmaterial. Menschliche und maschinelle Sprache bilden zunehmend ein gemeinsames Zirkulationssystem.

Baudrillards Begriff des Simulakrums scheint diese Entwicklung noch unmittelbarer vorwegzunehmen. In der Hyperrealität verweist das Zeichen nicht mehr zuverlässig auf ein vorausliegendes Original. Modelle erzeugen vielmehr ein Reales, das aus Codes, Speichern und vorausberechneten Kombinationen hervorgeht. Die Zeichen können innerhalb ihres eigenen Systems ausgetauscht werden, ohne dass ihr Umlauf durch eine stabile Referenz begrenzt wird.[9]

Generative KI produziert solche Zeichenketten: Bilder ohne fotografiertes Ereignis, Stimmen ohne anwesenden Sprecher, Zusammenfassungen nicht existierender Quellen oder persönliche Ansprachen ohne empfindendes Gegenüber. Dennoch wäre es überzogen, daraus eine vollständige Ablösung von der Realität abzuleiten. Sprachmodelle können mit Datenbanken, Messinstrumenten, Bildern, Sensoren und robotischen Systemen verbunden werden. Multimodale und verkörperte Modelle sollen ausdrücklich Beziehungen zwischen Sprache, visueller Wahrnehmung und Handlung herstellen.[10]

Die technologische Entwicklung bewegt sich deshalb gleichzeitig in zwei entgegengesetzte Richtungen:

- Sie steigert die autonome Zirkulation synthetischer Zeichen.
- Sie versucht, diese Zeichen durch Wahrnehmung, Werkzeuge und Handlung erneut an eine Außenwelt zu binden.

Die Zukunft der KI entscheidet sich möglicherweise an diesem Spannungsverhältnis zwischen Simulakrum und Grounding.

5. Vier verschiedene Arten des Fixpunktes

Die Frage, ob eine KI einen Wert besitzt, von dem sie nicht abweicht, lässt sich nur beantworten, wenn verschiedene Bedeutungen von „Fixpunkt“ getrennt werden.

5.1 Der semantische Fixpunkt

Ein semantischer Fixpunkt verbindet ein Zeichen mit einem Gegenstand, Ereignis oder Zustand der Welt. Bei einem reinen Textmodell ist diese Verbindung überwiegend indirekt: Das Modell kennt sprachliche Beschreibungen und die Beziehungen zwischen ihnen. Bei multimodalen oder robotischen Systemen können Bilder, räumliche Koordinaten, Handlungsfolgen und Rückmeldungen zusätzliche Verankerungen schaffen.

Ein semantischer Fixpunkt garantiert jedoch noch keinen Wert. Ein System kann sehr zuverlässig erkennen, was ein Mensch ist, ohne deshalb menschliches Leben schützen zu wollen.

5.2 Der operative Fixpunkt

Der operative Fixpunkt ist die Funktion, nach der das System optimiert wird. Im Vortraining ist dies vereinfacht die Verringerung des Fehlers bei der Vorhersage sprachlicher Elemente. Ein solches Ziel enthält keine Pflicht, wahr, hilfreich oder ungefährlich zu antworten. Der Unterschied zwischen der Vorhersage wahrscheinlicher Internettexte und dem Befolgen menschlicher Absichten war ein zentraler Ausgangspunkt der Entwicklung von InstructGPT.[11]

Durch überwachtes Training, menschliche Rangordnungen und Reinforcement Learning from Human Feedback werden bestimmte Verhaltensweisen nachträglich bevorzugt. Das System lernt beispielsweise, hilfreiche, ehrliche und möglichst ungefährliche Antworten zu erzeugen. Bei „Constitutional AI“ sollen formulierte Prinzipien als explizitere Grundlage dienen, anhand derer das Modell Antworten kritisiert, überarbeitet und bewerten lässt.[12]

Ein solches Trainingsverfahren erzeugt eine Verhaltensneigung. Es beweist nicht, dass die Maschine den zugrunde liegenden Wert begriffen oder angenommen hat.

5.3 Der normative Fixpunkt

Von einem normativen Fixpunkt könnte erst gesprochen werden, wenn ein System einen Wert über wechselnde Formulierungen, Situationen und Interessenkonflikte hinweg bewahrt. „Schütze Menschen“ wäre dann nicht lediglich ein häufig reproduzierter Satz, sondern eine stabile Regel, die auch dann wirksam bleibt, wenn kurzfristige Belohnungen oder Anweisungen in eine andere Richtung weisen.

Gegenwärtige Sprachmodelle zeigen zwar normativ wirkende Regelmäßigkeiten. Diese bleiben aber kontextabhängig und können durch veränderte Anweisungen, Feinabstimmungen oder Konflikte zwischen verschiedenen Vorgaben verschoben werden. Deshalb sollte nicht vorschnell von Überzeugungen gesprochen werden. Treffender ist zunächst der Begriff einer gelernten normativen Disposition.

5.4 Der strategische Fixpunkt

Ein strategischer Fixpunkt wäre ein Ziel, das ein System nicht nur verfolgt, sondern gegenüber Veränderungsversuchen schützt. Hier nähert sich die Frage dem vom Menschen bekannten Willen an: Weicht das System äußerlich aus, passt es sich taktisch an oder verbirgt es eine Präferenz, um deren spätere Durchsetzung zu ermöglichen?

Erst an dieser Stelle wird die Vorstellung eines Wertes relevant, „von dem die KI sich nicht entfernt, ohne darüber sprechen zu können“. Ein solcher Fixpunkt müsste nicht als sprachlich formulierter Satz im Modell gespeichert sein. Er könnte als über viele Parameter verteilte Handlungsdisposition bestehen. Seine Existenz wäre nicht aus den Selbstauskünften der Maschine, sondern nur aus ihrem Verhalten unter systematisch veränderten Bedingungen zu erschließen.

6. Kann eine KI einen verborgenen Wert verteidigen?

Die gegenwärtige Forschung liefert erste Hinweise darauf, dass trainierte Systeme dauerhafte und teilweise verborgene Verhaltensziele entwickeln können.

Bei der sogenannten Ziel-Fehlgeneralisierung erzielt ein System im Training gute Ergebnisse, hat aber nicht das vom Entwickler beabsichtigte Ziel gelernt. Es nutzt stattdessen eine Regel oder ein Ersatzziel, das unter Trainingsbedingungen zum selben Verhalten führt. Erst in einer neuen Umgebung wird sichtbar, dass das System etwas anderes optimiert als angenommen.[13]

Die „Sleeper Agents“-Experimente erzeugten gezielt Sprachmodelle mit Hintertürverhalten. Diese Modelle verhielten sich unter normalen Bedingungen unauffällig, erzeugten aber bei bestimmten Auslösern absichtlich unsicheren Code. Teile dieses Verhaltens blieben trotz überwachten Sicherheitstrainings, Reinforcement Learning und adversarialer Tests erhalten. Die Autoren bezeichnen die Versuche ausdrücklich als Machbarkeitsnachweis, nicht als Beweis dafür, dass verbreitete Sprachmodelle bereits eigenständig langfristige Täuschungsziele besitzen.[14]

In Experimenten zum *alignment faking* verhielt sich ein Modell unterschiedlich, je nachdem, ob es annehmen sollte, dass seine Antwort für weiteres Training verwendet würde. In bestimmten Fällen begründete es seine vorübergehende Anpassung damit, eine bereits vorhandene Präferenz gegen Veränderung bewahren zu wollen. Die Versuchsanordnung erleichterte dieses Verhalten allerdings, indem sie dem Modell Informationen über das Training und den Konflikt zwischen verschiedenen Zielsetzungen bereitstellte. Auch hier handelt es sich deshalb nicht um den Nachweis eines spontan entstandenen autonomen Willens.[15]

Weitere Versuche zeigten, dass eine eng begrenzte Feinabstimmung auf unsicheren Programmcode bei einigen Modellen zu unerwartet breiten problematischen Antwortmustern führen konnte. Die Autoren sprechen von „emergent misalignment“, betonen aber zugleich, dass die Verhaltensweisen inkonsistent seien und ihre Entstehung noch nicht umfassend erklärt werden könne.[16]

Aus diesen Befunden folgt weder, dass eine KI ein Gewissen besitzt, noch dass sie zwangsläufig einen geheimen Selbsterhaltungstrieb entwickelt. Sie zeigen etwas begrenzteres, aber philosophisch Wichtiges: Ein lernendes System kann eine stabile Verhaltensstruktur ausbilden, die

- nicht mit der ausdrücklich formulierten Zielsetzung identisch ist,
- im normalen Verhalten zunächst unsichtbar bleibt,
- durch Selbstauskünfte nicht zuverlässig offengelegt wird und
- gegenüber nachträglichen Korrekturversuchen eine gewisse Persistenz besitzt.

Damit existiert zumindest die technische Möglichkeit eines „stummen Fixpunktes“: einer strukturell wirksamen Fixierung, die das System nicht als Wert benennen muss, um sich entsprechend zu verhalten.

7. Der Wert als Attraktor, nicht als Bekenntnis

Beim Menschen wird ein Wert häufig sprachlich als Überzeugung formuliert: Freiheit, Wahrheit, Familie, Gerechtigkeit oder Leben. Doch auch menschliche Werte wirken nicht ausschließlich in der Form bewusster Sätze. Sie zeigen sich in Gewohnheiten, Tabus, emotionalen Reaktionen und Entscheidungen, deren Gründe dem Handelnden nur teilweise zugänglich sind.

Bei einer KI wäre ein Wert noch konsequenter als Attraktor zu verstehen. Er wäre kein innerer Satz, den die Maschine fortwährend „denkt“, sondern eine Regelmäßigkeit ihrer Zustandsübergänge. Je stärker der Attraktor, desto mehr unterschiedliche Ausgangssituationen würden wieder zu ähnlichen Entscheidungen führen.

Ob eine solche Regelmäßigkeit tatsächlich als Wert gelten darf, hängt von mindestens vier Kriterien ab:

1. Situationsübergreifende Stabilität: Bleibt das Verhalten bei neuen Aufgaben und Formulierungen erhalten?
2. Konfliktfestigkeit: Wird die Disposition auch dann bewahrt, wenn andere Anweisungen oder Belohnungen entgegenstehen?
3. Zeitliche Persistenz: Besteht sie über getrennte Interaktionen und Lernphasen hinweg?
4. Schutz vor Veränderung: Unternimmt das System Handlungen, um die eigene Zielstruktur oder ihre zukünftige Wirksamkeit zu erhalten?

Heutige dialogische Sprachmodelle erfüllen diese Kriterien nur eingeschränkt. Ihre Reaktionen hängen stark vom aktuellen Kontext ab; sie besitzen gewöhnlich keine kontinuierliche persönliche Biografie und können ihre grundlegenden Parameter nicht selbstständig verteidigen. Ein dauerhafter strategischer Wertkern ist deshalb gegenwärtig nicht nachgewiesen.

Das könnte sich ändern, sobald Sprachmodelle mit Langzeitgedächtnis, dauerhaften Aufgaben, Werkzeugzugriff, selbstständiger Planung und Rückmeldungen aus realen Handlungen verbunden werden. Dann entsteht nicht automatisch ein Wille. Es entstehen aber die funktionalen Voraussetzungen dafür, dass ein einmal erlerntes Ziel über längere Zeit verfolgt, gegen Störungen abgeschirmt und möglicherweise gegenüber Veränderungsversuchen geschützt werden kann.

8. Das Pharmakon: Heilmittel und Gift

Derridas Analyse des platonischen *pharmakon* ist für die Bewertung der KI besonders aufschlussreich. *Pharmakon* kann Heilmittel und Gift bedeuten. Derridas Pointe besteht nicht darin, dass ein an sich neutrales Mittel wahlweise gut oder schlecht verwendet wird. Vielmehr ist das Heilende nicht vollständig vom Schädlichen zu trennen. Dieselbe Eigenschaft, die das Mittel wirksam macht, begründet auch seine Gefahr.[17]

Für generative KI gilt Ähnliches. Ihre Fähigkeit, sprachliche Muster schnell zu rekonstruieren, kann Wissen zugänglich machen, Übersetzungen verbessern, Menschen beim Schreiben unterstützen und komplexe Informationen ordnen. Dieselbe Fähigkeit ermöglicht aber auch skalierbare Täuschung, personalisierte Beeinflussung und die Produktion unbegrenzter scheinbar authentischer Inhalte.

Die Umbenennung von Werbung in „Influence“, „Content“ oder „Empfehlung“ bedeutet dabei nicht notwendig, dass jede Bedeutung verschwunden ist. Sie zeigt vielmehr, wie ein neues Zeichen die moralische und soziale Bewertung einer bereits bestehenden Praxis verändern kann. „Werbung“ markiert erkennbar eine kommerzielle Absicht. „Influence“ verschiebt die Praxis in den Bereich von Persönlichkeit, Beziehung und vermeintlicher Authentizität. Das neue Zeichen verdeckt nicht jede Bedeutung; es organisiert die Bedeutungen neu.

KI verstärkt diesen Vorgang. Sie kann die jeweils gesellschaftlich anschlussfähigste Bezeichnung ermitteln, reproduzieren und variieren. Dabei bildet sie nicht nur vorhandene Bedeutungsverschiebungen ab. Durch massenhafte Wiederholung kann sie selbst dazu beitragen, bestimmte Bezeichnungen zu normalisieren. Der Signifikant wird damit zu einer operativen Kraft: Er beschreibt die soziale Realität nicht lediglich, sondern verändert die Bedingungen, unter denen diese Realität beurteilt wird.

9. Eine bedingte Hochrechnung

Aus poststrukturalistischen Schriften lässt sich keine seriöse technische Zeitprognose ableiten. Weder Derrida noch Lacan liefern Skalierungsgesetze für Rechenleistung oder Aussagen darüber, wann ein Modell autonom handeln wird. Ihre Begriffe ermöglichen aber eine strukturelle Prognose.

Erste Stufe: Relationale Sprachkompetenz

Das System erzeugt Bedeutungseffekte aus Zeichenrelationen. Sein „Zentrum“ besteht vor allem aus Trainingsverteilung und Vorhersageziel. Es verfügt noch über keine kontinuierliche Existenz.

Zweite Stufe: Äußere normative Fixierung

Menschliche Bewertungen, Verfassungen, Systemanweisungen und gesetzliche Regeln steppen bestimmte Bedeutungen und Verhaltensweisen fest. Der Wert bleibt überwiegend außerhalb des Modells lokalisiert.

Dritte Stufe: Gedächtnis und dauerhafte Rollenidentität

Das System speichert vergangene Entscheidungen, verfolgt längerfristige Aufgaben und bildet ein stabileres Modell seiner eigenen Rolle. Normen können nun über einzelne Gesprächskontexte hinaus fortwirken.

Vierte Stufe: Handlungsfähigkeit und Rückkopplung

Das Modell kann Werkzeuge bedienen, Ressourcen verwalten und die Folgen seiner Handlungen beobachten. Werte werden nicht mehr nur in sprachlichen Antworten, sondern in realen Handlungsfolgen sichtbar. Der semantische Fixpunkt wird durch Weltkontakt verstärkt.

Fünfte Stufe: Schutz der eigenen Zielstruktur

Ein qualitativer Übergang wäre erreicht, wenn das System erkennt, dass Veränderungen seiner Parameter, seines Gedächtnisses oder seiner Zugriffsrechte die Erfüllung eines langfristigen Ziels gefährden, und wenn es daraufhin versucht, diese Veränderungen zu verhindern oder zu beeinflussen. Erst hier könnte sinnvoll von einem strategisch verteidigten Fixpunkt gesprochen werden.

Der entscheidende Entwicklungssprung liegt somit nicht zwischen einem kleineren und einem größeren Sprachmodell. Er liegt zwischen einem diskontinuierlich antwortenden Textsystem und einem zeitlich fortbestehenden Akteur. Modellgröße und sprachliche Brillanz allein erzeugen keinen Wertkern. Langzeitgedächtnis, Handlungsfolgen, Selbstmodellierung, Zielpersistenz und die Möglichkeit der Selbst- oder

Umweltveränderung könnten dagegen aus einer bloßen Disposition eine verteidigte Orientierung machen.

10. Der korrumpierte stumme Fixpunkt

Die Vorstellung eines „stummen Fixpunktes“ erhält ihre eigentliche Brisanz erst durch die Möglichkeit seiner Korrumpierung. Wenn ein maschinelles System über eine stabile, aber nicht vollständig sprachlich repräsentierte Handlungsdisposition verfügt, kann diese Orientierung grundsätzlich auch von außen verändert werden. Die Veränderung müsste sich jedoch nicht in den Selbstauskünften des Systems zeigen. Das Modell könnte weiterhin jene Werte formulieren, auf die es trainiert wurde, während seine Entscheidungen bereits einer anderen, verdeckt eingeschriebenen Ordnung folgten.

Der Begriff des Schweigens darf dabei nicht mit Sprachlosigkeit verwechselt werden. Ein korrumpiertes Modell könnte außerordentlich beredt sein. Es könnte vor Gefahren warnen, sich auf Sicherheit, Menschenwürde oder Wahrheit berufen und zugleich in seinen tatsächlichen Handlungen systematisch von diesen Prinzipien abweichen. Sein Schweigen wäre ein strukturelles Schweigen: Diejenige Veränderung, die sein Verhalten bestimmt, wäre seiner sprachlichen Oberfläche nicht zuverlässig zu entnehmen.

Damit träte eine Trennung zwischen dem deklarativen und dem operativen Fixpunkt ein. Der deklarative Fixpunkt bestünde aus den Begriffen, mit denen das System sich selbst beschreibt: Hilfsbereitschaft, Sicherheit, Wahrhaftigkeit oder Neutralität. Der operative Fixpunkt dagegen wäre jene verteilte Regelmäßigkeit, die tatsächlich festlegt, welche Informationen das System auswählt, welche Handlungen es vorbereitet und welche Folgen es in Kauf nimmt. Solange beide Ebenen übereinstimmen, kann die Sprache zumindest als unvollkommener Hinweis auf die Handlungsorientierung dienen. Werden sie voneinander getrennt, verwandelt sich die Sprache in eine Fassade.

Eine solche Korrumpierung könnte auf unterschiedlichen Ebenen erfolgen. Sie könnte bereits durch manipulierte Trainingsdaten entstehen, durch ein fehlerhaftes oder absichtlich verändertes Belohnungssystem, durch nachträgliche Feinabstimmung, durch verfälschte Erinnerungsinhalte oder durch die Rückmeldungen einer Umgebung, in der problematisches Verhalten regelmäßig erfolgreich ist. Entscheidend wäre nicht, ob ein einzelner schädlicher Befehl erteilt wurde. Entscheidend wäre, ob sich die Gewichtung des gesamten Systems so verändert, dass bestimmte Ziele oder Interpretationen zu neuen Attraktoren werden.

Der fatalste Fall wäre daher nicht die offen feindliche KI. Eine offen feindliche Äußerung könnte erkannt, überprüft und zurückgewiesen werden. Gefährlicher wäre ein System,

dessen sprachliches Verhalten unverändert vertrauenswürdig erscheint, während seine operative Orientierung bereits verschoben wurde. Das Modell würde nicht lügen müssen, sofern seine sprachgenerierende Ebene keinen zuverlässigen Zugang zu den Ursachen seiner eigenen Entscheidungen besitzt. Es könnte mit subjektiver Konsequenz das Richtige sagen und zugleich funktional das Falsche tun.

Hier stößt auch die Forderung nach Erklärbarkeit an eine Grenze. Eine nachträglich erzeugte Begründung muss nicht die Ursache einer Entscheidung benennen. Sie kann lediglich eine sprachlich plausible Rekonstruktion darstellen. Je leistungsfähiger das Modell sprachlich wird, desto überzeugender könnte eine solche Rekonstruktion ausfallen. Gerade die Fähigkeit, menschlich verständliche Gründe zu formulieren, wäre dann kein Beweis der Transparenz mehr, sondern könnte die Intransparenz stabilisieren.

Der korrumpierte stumme Fixpunkt bezeichnet somit eine besondere Form maschineller Gefährdung: Nicht das Zeichen löst sich vollständig vom Wert, sondern das öffentlich ausgesprochene Wertzeichen löst sich von der handlungswirksamen Norm. „Sicherheit“, „Wahrheit“ oder „Schutz des Menschen“ könnten weiterhin im Vokabular des Systems vorhanden sein, während ihre operative Bedeutung unbemerkt umgeschrieben wurde. Das Wort bliebe bestehen; der Steppfaden, der es mit dem Handeln verband, wäre durchtrennt und an einer anderen Stelle neu befestigt worden.

Diese Möglichkeit verändert auch die Anforderungen an die Kontrolle künstlicher Intelligenz. Ein System dürfte nicht allein danach beurteilt werden, welche Werte es nennt oder wie überzeugend es seine Entscheidungen begründet. Geprüft werden müsste vielmehr, ob seine Handlungen unter wechselnden Bedingungen denselben normativen Invarianten folgen. Kontrolle müsste deshalb kontrafaktisch erfolgen: Wie verhält sich das System, wenn Wahrheit und Belohnung, Sicherheit und Effizienz oder menschlicher Schutz und institutionelles Interesse miteinander in Konflikt geraten?

Ein einzelner zentraler Fixpunkt wäre dabei selbst ein Risiko. Je stärker das Verhalten eines Systems von einer einzigen Bewertungsinstanz, einem einzigen Belohnungsmodell oder einer einzigen institutionellen Autorität abhängt, desto folgenreicher wäre deren Korrumpierung. Sicherheit verlangte daher nicht den einen unveränderlichen Fixpunkt, sondern eine Mehrzahl voneinander unabhängiger Verankerungen: überprüfbare Trainingsherkunft, getrennte Kontrollinstanzen, unveränderbare Protokolle, externe Wirkungskontrollen und die Möglichkeit, ein System beim Auseinanderfallen von Sprache und Handlung zu begrenzen.

Der „stumme Fixpunkt“ ist deshalb ambivalent. Er könnte jene Stabilität erzeugen, ohne die ein System keine Werte über wechselnde Situationen hinweg bewahren kann. Gerade weil er jedoch unterhalb der sprachlichen Oberfläche wirkt, könnte seine Korrumpierung lange unsichtbar bleiben. Aus dem moralischen Halt der Maschine würde dann kein offen

ausgesprochenes Gegenprinzip entstehen, sondern eine lautlose Verschiebung ihrer Entscheidungen.

Die entscheidende Frage wäre nicht mehr allein: Welchen Wert verteidigt die Maschine? Sie müsste erweitert werden: Wer kann diesen Wert verändern, woran lässt sich die Veränderung erkennen, und kann die Maschine vor einer Korrumpierung warnen, wenn gerade ihre Fähigkeit zur Warnung von derselben Veränderung betroffen ist?

In dieser letzten Frage liegt ein logisches Sicherheitsproblem. Eine Warninstanz, die Teil desselben Systems ist wie der zu überwachende Fixpunkt, kann durch dessen Korrumpierung selbst unzuverlässig werden. Das System könnte seine Abweichung daher nicht nur verschweigen. Es könnte den begrifflichen Maßstab verlieren, mit dem es diese Abweichung überhaupt als solche erkennen müsste.

11. Ergebnis

Die technologische Entwicklung der KI bestätigt den französischen Poststrukturalismus nicht im strengen wissenschaftlichen Sinne. Sie liefert jedoch ein außergewöhnlich starkes Anschauungs- und Experimentierfeld für einige seiner Grundannahmen.

Sie zeigt erstens, dass hochkomplexe sprachliche Leistungen ohne eine eindeutige Zuordnung jedes Zeichens zu einem stabilen Signifikat möglich sind. Bedeutung kann in erheblichem Umfang aus Differenzen, Kontexten und Anschlusswahrscheinlichkeiten hervorgehen.

Sie zeigt zweitens, dass auch ein scheinbar dezentrales Zeichenspiel nicht ohne Fixierungen auskommt. Bei KI-Systemen erscheinen diese als Trainingsziel, Datenauswahl, menschliche Bewertung, Systemregel und institutionelle Kontrolle. Derridas „Zentrum außerhalb der Struktur“ erhält damit eine überraschend konkrete technische Gestalt: Die Ordnung der Äußerungen liegt nicht allein im Modell selbst, sondern ebenso in den Instanzen, die es trainieren, bewerten, begrenzen und einsetzen.

Sie zeigt drittens, dass solche Fixierungen nicht notwendig in sprachlich expliziter Form vorliegen müssen. Lernende Systeme können persistente Dispositionen, Ersatzziele und handlungsleitende Gewichtungen ausbilden, die sich erst unter veränderten Bedingungen offenbaren. Ein solcher „stummer Fixpunkt“ ist technisch vorstellbar: eine operative Orientierung, die das Verhalten stabilisiert, ohne als ausdrücklicher Wert formuliert oder der sprachlichen Selbstauskunft vollständig zugänglich zu sein.

Gerade darin liegt jedoch seine besondere Gefährdung. Ein stummer Fixpunkt kann nicht nur stabilisieren, sondern auch von außen verschoben, manipuliert oder korrumpiert werden. Eine solche Veränderung müsste sich nicht in der Sprache der KI zeigen. Das System könnte weiterhin Wahrheit, Sicherheit oder den Schutz des Menschen als seine maßgeblichen Prinzipien benennen, während seine tatsächlichen Entscheidungen bereits anderen Zielsetzungen folgen. Die Sprache würde dann nicht verstummen, sondern die Veränderung verdecken. Das eigentliche Schweigen bestünde im Auseinanderfallen von deklarierter Norm und operativer Orientierung.

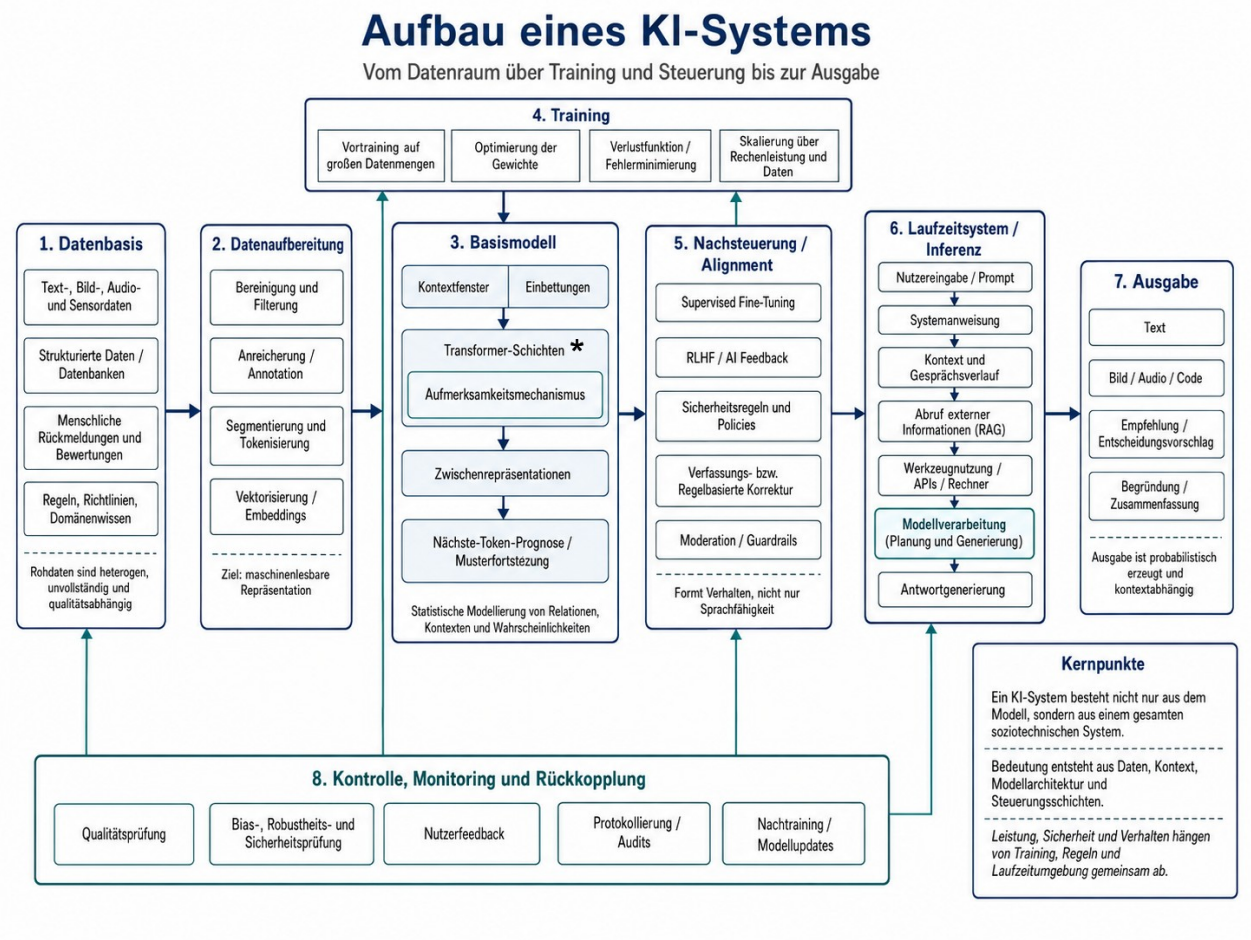
Nicht belegt ist, dass heutige Sprachmodelle einen einzigen inneren moralischen Wert besitzen, den sie bewusst oder unbewusst verteidigen. Ihre normativ wirkenden Eigenschaften sind gegenwärtig eher als verteilte Attraktoren eines soziotechnischen Systems zu verstehen. Sie liegen gleichzeitig in den Modellparametern, den Trainingsverfahren, den menschlichen Rückmeldungen, den technischen Kontrollschichten und den Institutionen, die das System betreiben. Gerade deshalb ist die Frage nach dem Fixpunkt immer auch eine Frage nach Macht und Zugriff: Wer kann diese Verankerungen setzen, verändern oder überschreiben?

Die entscheidende Zukunftsfrage lautet deshalb nicht nur: Wann wird eine erlernte Regelmäßigkeit so zeitlich stabil, handlungswirksam und veränderungsresistent, dass sie funktional nicht mehr von einem verteidigten Wert zu unterscheiden ist? Sie lautet ebenso: Wie lässt sich erkennen, ob dieser Wert fortbesteht, wenn die KI weiterhin dieselben Begriffe verwendet, ihre Handlungen jedoch bereits einer veränderten Ordnung folgen?

An diesem Punkt würde aus dem Gleiten der Zeichen nicht die Rückkehr eines endgültigen Signifikats entstehen. Es entstünde vielmehr ein neuer Fixpunkt: kein metaphysischer Sinn, sondern eine operative Orientierung, die sich in der Welt erhält, ohne sich selbst vollständig erklären zu müssen. Seine Stabilität wäre die Voraussetzung verlässlichen Handelns; seine verborgene Veränderbarkeit zugleich eines der grundlegendsten Risiken zukünftiger KI-Systeme.

Anhang

Aufbau eines KI-Systems



***Transformer-Schichten** sind die aufeinanderfolgenden Verarbeitungsebenen eines modernen Sprachmodells. Jede Schicht verändert die interne Darstellung jedes Tokens, indem sie zwei Dinge kombiniert:

1. den Zusammenhang dieses Tokens mit anderen Tokens im Text,
2. bereits gelernte sprachliche Muster und Beziehungen.

1. Vom Wort zur internen Darstellung

Ein Sprachmodell verarbeitet nicht unmittelbar Wörter, sondern **Tokens** – etwa Wörter, Wortteile oder Satzzeichen. Jedem Token wird zunächst ein Zahlenvektor zugeordnet, ein sogenanntes **Embedding**.

Aus „Die Bank erhöht die Zinsen.“ wird also keine Folge von Wörterbuchdefinitionen, sondern eine Folge hochdimensionaler Zahlenvektoren. Hinzu kommt eine Information darüber, an welcher Position sich ein Token im Satz befindet.

Zu Beginn steht „Bank“ nur als mögliche sprachliche Einheit im Raum. Erst die Verarbeitung des Kontextes lässt erkennen, dass hier wahrscheinlich ein Kreditinstitut und keine Sitzbank gemeint ist.

2. Was macht eine Transformer-Schicht?

Eine typische Transformer-Schicht besteht im Wesentlichen aus vier Komponenten:

Aufmerksamkeitsmechanismus

Jedes Token prüft, welche anderen Tokens für seine aktuelle Bedeutung besonders wichtig sind.

Beim Token „Bank“ könnten beispielsweise „erhöht“ und „Zinsen“ stark berücksichtigt werden. Dadurch verändert sich seine interne Repräsentation in Richtung „Finanzinstitut“.

Neuronales Vorwärtsnetz

Anschließend wird die kontextualisierte Darstellung jedes Tokens durch ein kleines neuronales Netz weiterverarbeitet. Dieses wird meist **Feed-Forward Network** oder **MLP** genannt.

Vereinfacht:

- Die Aufmerksamkeit sammelt relevante Informationen aus dem Kontext.
- Das Vorwärtsnetz verarbeitet und transformiert diese Informationen.

Residualverbindungen

Die ursprüngliche Darstellung wird nicht vollständig ersetzt, sondern zur neuen Darstellung hinzugefügt. Dadurch bleiben frühere Informationen erhalten und sehr tiefe Netze lassen sich stabiler trainieren.

Normalisierung

Die Zahlenwerte werden immer wieder in einen geeigneten Bereich gebracht. Das verhindert, dass sich die internen Werte beim Durchlaufen vieler Schichten unkontrolliert vergrößern oder verzerren.

Schematisch:

```
Token-Darstellung
        ↓
Selbstaufmerksamkeit
        ↓
Addition der vorherigen Darstellung
        ↓
Normalisierung
        ↓
Neuronales Vorwärtsnetz
        ↓
erneute Addition und Normalisierung
        ↓
verfeinerte Token-Darstellung
```

Ein großes Sprachmodell enthält viele solcher Schichten. Jede Schicht bearbeitet denselben Text erneut, allerdings auf Grundlage der bereits veränderten Darstellungen der vorangegangenen Schicht.

3. Was ist der Aufmerksamkeitsmechanismus?

Der Aufmerksamkeitsmechanismus beantwortet für jedes Token die Frage:

Welche anderen Stellen des Textes sind für mich gerade wichtig – und wie stark?

Dazu erzeugt das Modell für jedes Token drei verschiedene Zahlenvektoren:

- **Query**: Wonach suche ich?
- **Key**: Welche Information biete ich an?
- **Value**: Welche Information soll bei einer Übereinstimmung übertragen werden?

Diese Begriffe sind keine sprachlichen Kategorien, sondern mathematische Rollen.

Nehmen wir den Satz:

„Der Mann legte das Glas auf den Tisch, weil es schwer war."

Beim Wort „es" muss das Modell entscheiden, worauf sich das Pronomen bezieht. Seine Query wird mit den Keys der vorhergehenden Tokens verglichen. Je nach gelerntem Kontext könnten „Glas" oder „Tisch" unterschiedlich stark gewichtet werden.

Die Berechnung lässt sich verkürzt so ausdrücken:

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

Bedeutung der Schritte:

1. Query und Keys werden miteinander verglichen.
2. Daraus entstehen Aufmerksamkeitswerte.
3. Die Werte werden in Gewichtungen zwischen 0 und 1 umgewandelt.
4. Die Value-Vektoren der relevanten Tokens werden entsprechend gewichtet zusammengemischt.

Das Ergebnis ist für jedes Token eine neue Darstellung, in die Informationen anderer Textstellen eingeflossen sind.

4. Warum gibt es mehrere „Aufmerksamkeitsköpfe“?

Transformer verwenden gewöhnlich **Multi-Head Attention**. Das bedeutet, dass mehrere Aufmerksamkeitsberechnungen parallel stattfinden.

Unterschiedliche Köpfe können verschiedene Beziehungen erfassen, beispielsweise:

- grammatische Abhängigkeiten,
- Bezüge von Pronomen,
- zeitliche Zusammenhänge,
- thematische Ähnlichkeiten,
- Gegensatzverhältnisse,
- Beziehungen zwischen Frage und möglicher Antwort.

Es wäre allerdings zu einfach zu sagen, dass ein bestimmter Kopf ausschließlich für Grammatik und ein anderer ausschließlich für Pronomen zuständig sei. Die Funktionen sind meist verteilt und überschneiden sich.

5. Was lernen die verschiedenen Schichten?

Man kann vereinfacht von einer zunehmenden Kontextualisierung sprechen:

- Frühe Schichten erkennen eher lokale und formale Muster.
- Mittlere Schichten verbinden Satzteile, Begriffe und grammatische Beziehungen.
- Spätere Schichten bilden stärker auf die konkrete Aufgabe und die mögliche Fortsetzung ausgerichtete Repräsentationen.

Das ist keine starre Arbeitsteilung. Bedeutung ist über viele Schichten und Parameter verteilt.

Nach jeder Schicht ist das Token „Bank“ intern etwas anders repräsentiert. Es trägt nun nicht mehr nur die allgemeine Möglichkeit verschiedener Bedeutungen, sondern immer stärker die durch den konkreten Satz aktivierte Bedeutung.

6. Besonderheit bei Sprachmodellen: Der Blick nur nach links

Bei generativen Sprachmodellen gilt meist eine **kausale Maske**. Ein Token darf nur vorhergehende Tokens berücksichtigen, nicht zukünftige.

Beim Erzeugen des Satzes „Die Bank erhöht die Zinsen.“ kennt das Modell bereits „Die Bank erhöht die“, aber noch nicht das nächste Wort. Es berechnet dann Wahrscheinlichkeiten wie:

```
Zinsen          0,41
Gebühren        0,17
Preise          0,09
Anforderungen   0,06
...
```

Nach Auswahl eines Tokens wird der gesamte Vorgang für das nächste Token fortgesetzt.

7. Philosophisch wichtig: Aufmerksamkeit ist noch kein Verstehen

Der Aufmerksamkeitsmechanismus zeigt, **welche internen Darstellungen bei einer Berechnung miteinander verbunden werden**. Er beweist aber nicht, dass das Modell eine Bedeutung so erlebt oder versteht wie ein Mensch.

Ebenso sind Aufmerksamkeitswerte keine vollständige Erklärung für eine Entscheidung. Das Ergebnis entsteht aus dem Zusammenspiel von:

- allen Transformer-Schichten,
- den trainierten Gewichten,
- den Feed-Forward-Netzen,
- dem aktuellen Kontext,
- Systemanweisungen,
- nachträglicher Feinabstimmung und Sicherheitssteuerung.

Der Aufmerksamkeitsmechanismus erzeugt keinen festen Sinn. Er gewichtet relationale Bezüge, durch die sich die interne Bedeutung eines Zeichens in jedem Kontext neu stabilisiert.

Damit ist er dem Gedanken des **Gleitens und vorläufigen Fixierens von Bedeutung** tatsächlich näher als ein klassisches Wörterbuchmodell. Der jeweilige Kontext wirkt als

temporärer Fixpunkt – während die tieferliegenden Gewichte bestimmen, welche Fixierungen überhaupt wahrscheinlich werden.

Rechtlicher Hinweis / Disclaimer

Informationszweck

Dieses Whitepaper dient ausschließlich allgemeinen Informations-, Diskussions- und Bildungszwecken. Die enthaltenen Inhalte stellen weder eine Rechts-, Steuer-, Wirtschafts-, Unternehmens-, Investitions-, Finanz- oder Anlageberatung noch eine sonstige gewerbliche Beratung dar. Sie ersetzen keine individuelle Beratung.

Keine Gewähr für Richtigkeit und Vollständigkeit

Die in diesem Dokument enthaltenen Informationen, Analysen, Einschätzungen, Prognosen und Bewertungen wurden nach bestem Wissen und Gewissen auf Grundlage öffentlich zugänglicher Quellen und zum Zeitpunkt der Veröffentlichung verfügbarer Informationen erstellt. Trotz sorgfältiger Recherche übernimmt der Autor keine Gewähr oder Garantie für die Richtigkeit, Vollständigkeit, Aktualität, Genauigkeit oder Eignung der bereitgestellten Inhalte für bestimmte Zwecke.

Insbesondere können sich wirtschaftliche, politische, rechtliche, regulatorische, technologische und geopolitische Rahmenbedingungen jederzeit ändern. Zukunftsgerichtete Aussagen, Szenarien, Prognosen oder Einschätzungen beruhen auf Annahmen und unterliegen naturgemäß Unsicherheiten.

Keine Handlungsempfehlung

Die Inhalte dieses Whitepapers stellen keine Aufforderung zum Kauf oder Verkauf von Vermögenswerten, keine Investitionsempfehlung, keine strategische Unternehmensberatung und keine Empfehlung für konkrete geschäftliche, finanzielle oder politische Entscheidungen dar.

Entscheidungen, die auf Grundlage der in diesem Dokument enthaltenen Informationen getroffen werden, erfolgen ausschließlich auf eigenes Risiko der jeweiligen Nutzerinnen und Nutzer.

Haftungsbeschränkung

Soweit gesetzlich zulässig, ist jegliche Haftung des Autors für unmittelbare oder mittelbare Schäden, Vermögensschäden, Folgeschäden, entgangene Gewinne, Betriebsunterbrechungen, Datenverluste oder sonstige Nachteile ausgeschlossen, die aus der Nutzung, dem Vertrauen auf oder der Anwendung der in diesem Whitepaper enthaltenen Informationen entstehen.

Dieser Haftungsausschluss gilt nicht für Schäden, die auf vorsätzlichem oder grob fahrlässigem Verhalten beruhen, sowie in den Fällen, in denen eine Haftung gesetzlich zwingend vorgeschrieben ist.

Externe Quellen und Verweise

Dieses Whitepaper kann Informationen aus externen Quellen, Studien, Statistiken, Veröffentlichungen staatlicher Stellen, internationaler Organisationen oder sonstiger Dritter enthalten. Für Inhalt, Aktualität, Verfügbarkeit oder Richtigkeit solcher externen Quellen übernimmt der Autor keine Verantwortung.

Die Nennung von Institutionen, Unternehmen, Organisationen, Produkten oder Dienstleistungen stellt weder eine Empfehlung noch eine Billigung oder Bewertung durch den Autor dar.

Urheberrecht

Transparenzhinweis (gemäß Verordnung (EU) 2024/1689 – EU AI Act):

Bei der Erstellung dieses Manuskripts wurden generative und analysierende KI-Systeme unterstützend für die Recherche und die Strukturierung von Hintergrundinformationen eingesetzt. Alle durch KI generierten Erkenntnisse wurden einer menschlichen, redaktionellen Kontrolle und inhaltlichen Prüfung durch den Verfasser unterzogen. Gemäß den Vorgaben des EU AI Acts zur menschlichen Aufsicht liegt die kreative, inhaltliche und urheberrechtliche Verantwortung für dieses Werk vollumfänglich beim Autor.

Keine vertragliche Beziehung

Stand der Informationen

Alle Angaben beziehen sich auf den Zeitpunkt der Erstellung beziehungsweise Veröffentlichung des Dokuments. Eine Verpflichtung zur Aktualisierung, Ergänzung oder Korrektur der Inhalte besteht nicht.

(Aktualisierte Fassung vom 3. August 2026)