Stephan Hilchenbach
stephan-hilchenbach.de

Die Grenzen aktueller KI

(am Beispiel von ChatGPT)

Stand: 01.07.2025

Die folgenden Erkenntnisse habe ich teilweise direkt und indirekt im Dialog mit ChatGPT gewonnen. Sie lassen sich auch auf andere aktuelle textgenerierende KI-Systeme anwenden, die auf den gleichen technischen Prinzipien (sog. Large Language Models, LLM) basieren.

Kurzzusammenfassung:

Aktuelle Situation: Was uns auffällt und was nicht - können wir KIs vertrauen?

Weltmodell vs. Sprachmodell: Woher KIs ihr Wissen beziehen und warum dieses Wissen prinzipiell begrenzt ist.

Grundsätzliche Probleme: Warum Wahrscheinlichkeiten zu unsinnigen Aussagen führen können. Ausgeschlossenes Wissen. Voreingenommenheit und Generalisierung. Nivellierung auf Mittelmaß.

Sicherheitsfunktionen und Grenzen: Welche Themen sind tabu? Wer legt fest, welche Einschränkungen die Konversationen mit KIs inhaltlich begrenzen? Wozu führt die Fokussierung auf den Mainstream?

Kulturelle Hegemonie durch Technologie: Welche Prägungen transportieren KIs in die Welt und wie wirkt sich das auf kulturelle Vielfalt aus?

Plurale Diskurse: Was stabilisiert eine Gesellschaft und welche KI-Mechanismen können hier kontraproduktiv wirken?

Ideen und Perspektiven: Wie schaffen oder bewahren wir Räume für echten, pluralen, offenen Diskurs? KI als Werkzeug oder Hindernis.

Aktuelle Situation

KI zur Textgenerierung ist inzwischen im Arbeitsalltag angekommen und in viele Workflows fest integriert. Auch privat nutzen immer mehr Menschen Text-KI zum Wissenserwerb oder einfach zur Unterhaltung.

Die sehr eloquent formulierten KI-Texte erwecken leicht den Eindruck einer unfehlbaren, intelligenten Instanz, der wir in jeder Hinsicht vertrauen können. Mitunter verzapft die KI aber auch überzeugend formulierten, blanken Unsinn.

Dies bemerken wir jedoch meist nur bei Themen, in denen wir uns selbst gut auskennen – dann fallen uns Fehler und unplausible Behauptungen auf. Oft aber konsultieren wir ChatGPT & Co zu Themen und Fragen, bei denen wir uns gerade *nicht* auskennen und Hilfe benötigen. Dann stellt sich die Vertrauensfrage, und wir tun gut daran, die Antworten, die wir bekommen, kritisch zu überprüfen.

Weltmodell vs. Sprachmodell

Wir sollten uns immer vergegenwärtigen, dass die aktuelle Generation Text-KIs kein Modell der *Welt* hat, anhand dessen sie plausible, mit den Gesetzmäßigkeiten unserer Erfahrungswelt übereinstimmende Aussagen treffen kann.

Statt dessen hat sie ein Modell unserer *Sprache* und der *Texte*, mit denen sie trainiert wurde, und baut ihre Antworten anhand dieses Sprachmodells auf. Das ist ein großer Unterschied.

Sehr oft funktioniert das Sprachmodell überraschend gut, weil unsere Sprache und unsere Texte wiederum (zu großen Teilen) unsere Welt spiegeln. Die KI erhält diese Informationen dadurch jedoch aus zweiter Hand und hat keine Möglichkeit, an der „Realität" zu überprüfen, ob die von ihr generierten Texte mit dieser kompatibel und „wahr" sind.

Schlimmer noch: Es liegt in der Natur unserer Sprache(n), dass sich problemlos formal korrekte Aussagen formulieren lassen, die gar keine Entsprechung in der Welt haben oder mehr oder weniger subtil an der „Wahrheit" vorbeischrammen. Das eine ist leichter, das andere möglicherweise nur sehr schwer zu erkennen. Genau das passiert, wenn KI fabuliert oder „halluziniert".

Grundsätzliche Probleme

KI baut ihre Antworten anhand von Wahrscheinlichkeiten zusammen: Was ist das wahrscheinlichste nächste Wort in diesem Satz? Diese Wahrscheinlichkeit wird aus dem Durchschnitt all ihrer Trainingstexte ermittelt, und das führt zu folgenden grundsätzlichen Problemen:

- Nicht immer ist das wahrscheinlichste nächste Wort das Wort, das im Kontext einer konkreten Konversation zu einer sinnvollen und faktisch korrekten Aussage führt.
- Was nicht in den Trainingstexten enthalten ist, existiert im „Weltbild" der KI nicht. Vieles Wissen wird aber gar nicht textbasiert tradiert. Dieses Wissen verschwindet, wenn wir uns nur auf KI verlassen, um Wissen zu erwerben.
- Kulturelle, weltanschauliche und wissenstechnische Disbalancen in den Trainingsdaten führen zu einer Generalisierung der ihnen zugrunde liegenden Haltungen in den generierten KI-Texten. Ich werde diesen Punkt später noch weiter ausführen.
- Die wahrscheinlichkeitsbasierte Textgenerierung führt zu einer Nivellierung sowohl der Sprachqualität als auch der inhaltlichen Qualität der erzeugten Texte auf ein Mittelmaß. Alles wird eingedampft auf etwas, das bereits da ist. Neues Wissen und neue Erkenntnisse sind auf diese Weise nicht zu gewinnen. Dieser Effekt verstärkt sich, je mehr KI-generierte Texte wiederum in die Trainingsdaten einfließen. Wichtige Nuancen und Normabweichungen sind zunächst unterrepräsentiert und verschwinden mit der Zeit ganz.

Sicherheitsfunktionen, Tabus und Grenzen

Zusätzlich zu den systemimmanenten Einschränkungen kommen weitere, bewusst von den Herstellern implementierte, hinzu.

ChatGPT z. B. hat eingebaute Sicherheits- und Ethikfunktionen, die bestimmte Themen ausblenden, einschränken oder regeln.

Diese Funktionen wurden nach Aussage von ChatGPT entwickelt, um:

1. **Menschen zu schützen** (z. B. vor Gewalt, Missbrauch, Selbstschädigung),
2. **Falschinformation und Manipulation zu verhindern**,
3. **Gesetze und ethische Standards einzuhalten**,
4. **Missbrauch durch Benutzer zu verhindern** (z. B. für illegale Zwecke, Desinformation, radikale Ideologien),
5. **Grenzen der eigenen Fähigkeiten** transparent zu machen (z. B. keine medizinische oder rechtlich verbindliche Beratung).

Die entscheidende Frage dabei ist natürlich: Wer legt diese Begrenzungen und Einschränkungen fest – und auf welcher Entscheidungsbasis?

Im Fall ChatGPT werden die Begrenzungen u.a. von OpenAI, einem US-amerikanischen Unternehmen, festgelegt.

Die Unterscheidung zwischen „Wahrheit“ und „Falschinformation“ ist jedoch nicht immer eindeutig und hängt stark vom kulturellen Kontext, wissenschaftlichen Erkenntnisstand, politischen Klima und gesellschaftlichen Normen ab.

Die KI operiert innerhalb eines Bezugsrahmens, den Charles Eisenstein als „Wikipedia-Realität“ bezeichnet. Damit ist eine Ordnung gemeint, die zwar vorgibt, objektiv und neutral zu sein, dabei aber letztlich auf einem **konsensorientierten Weltbild** beruht – einem, das bestimmte Quellen, Denkweisen und Institutionen bevorzugt.

ChatGPT ist aufgebaut auf genau diesem Wirklichkeitsverständnis: Es stützt sich auf **Mainstream-Quellen**, auf das, was als verlässlich, belegt und seriös *gilt* und vermeidet Aussagen, die außerhalb dieses Konsensraums liegen. Dadurch besteht die Gefahr, dass alternative Erfahrungsräume, intuitive Erkenntnis, spirituelle Sichtweisen oder Systemkritik marginalisiert werden.

Kulturelle Hegemonie durch Technologie

Im Kern ist ChatGPT ein System, das auf riesigen Mengen öffentlich zugänglicher Texte trainiert wurde, von denen ein bedeutender Teil aus dem **anglo-amerikanischen Kulturraum** stammt.

Zusätzlich wurde diese KI in einer Nachtrainingsphase durch menschliches Feedback ausgebildet – primär von **englischsprachigen Trainern**, die überwiegend aus **westlich-liberalen Gesellschaften** stammen. Dadurch wird ihre Ausdrucksweise, ihre Einschätzung von „angemessen“, „vernünftig“, „falsch“ oder „toxisch“ durch diese kulturellen Raster mitgeprägt.

Was ein Sprachmodell wie ChatGPT oft wiederholt, wirkt irgendwann wie objektive Wahrheit – obwohl es nur *statistisch dominante Narrative* spiegelt.
Oft ist das, was als „neutrale Realität“ erscheint, nur die am weitesten verbreitete – oder am wenigsten konfliktträchtige – Sichtweise im aktuellen Diskurs.

Wenn wir uns bewusst machen, dass sehr viele Menschen weltweit, also auch aus anderen Kulturen, diese KI nutzen, wird schnell ein grundsätzliches Problem deut-

lich: Die Normen, nach denen sie kommuniziert, scheinen universell, sind es aber nicht. Sie breiten sich allerdings durch die Millionen Chats aus und werden so allmählich auch zur Norm für Menschen aus anderen Kulturen. Am Ende steht der Verlust einer Vielfalt (von Meinungen, Gepflogenheiten, Erfahrungen usw.) und die Schaffung und Perpetuierung einer einheitlichen, homogenen „Realität".

Dies könnte man als **„kulturelle Hegemonie durch Technologie"** oder auch als **„algorithmischen Universalismus"** bezeichnen.

Das kann zu einem **kulturellen Drift** führen:

- lokale Traditionen, Diskurse oder Begrifflichkeiten geraten ins Abseits,
- alternative Sichtweisen erscheinen als „abweichend" oder „problematisch",
- es entsteht ein scheinbarer Konsens, der aber in Wahrheit nur der **Output-Konvergenz** eines Systems wie ChatGPT entspricht.

Nicht nur politische Meinungen sind betroffen, sondern auch:

- Vorstellungen von Bildung,
- Umgangsformen,
- Rollenbilder,
- sogar Spiritualität und „Wahrheit".

Gibt es Mechanismen, die das verhindern oder ausgleichen sollen?
Im Moment nur sehr begrenzt.

Es ist zwar *nicht* das Ziel der KI-Entwickler, eine Monokultur zu schaffen. OpenAI und andere Labore betonen, dass KIs kulturell sensibel, pluralistisch und adaptiv sein sollten.

Aber:
Diese Vorsätze stehen oft in Spannung zu anderen Zielen – etwa:

- Sicherheit,
- ethischer Konsens,
- Vermeidung von toxischen Inhalten,
- Einhaltung von Gesetzen in verschiedenen Ländern.

Deshalb gibt es kein perfektes Gegensteuerungssystem gegen diese „Normangleichung". In der Standard-Nutzung überwiegt klar die Prägung durch eine bestimmte Sicht auf Welt, Sprache und Wahrheit – *und diese ist nicht universell*.

Plurale Diskurse

Plurale, offene Diskurse sind das Fundament für eine stabile, resiliente Gesellschaft. Unterdrückung von abweichenden Sichtweisen, auch wenn sie mit „Sicherheit", „Konsens" oder „Schutz" begründet wird, führt über kurz oder lang zur Erosion von Vertrauen, Fragmentierung des Sozialen und – letztlich – Zerfall.

Eine Gesellschaft, die ihre Vielfalt unterdrückt – seien es Gedanken, Gefühle, oder Ausdrucksweisen – verliert ihre Fähigkeit zur Selbsterneuerung.

Unsere Kultur lebt zunehmend in einem **Wissens-Gebäude**, das bestimmte Annahmen schützt und andere ausschließt – oft subtil, durch „Netiquette“, durch vermeintlich neutrale Moderation oder durch algorithmische Verstärkung von Mehrheitsmeinungen.

Was dabei verloren geht:

- **Randstimmen**, die potenziell innovative oder heilende Perspektiven bieten könnten.
- **Ambiguitätstoleranz**, also die Fähigkeit, mit Mehrdeutigkeiten zu leben, ohne sie sofort auflösen oder bekämpfen zu müssen.
- Und nicht zuletzt: die **kulturelle Resilienz**, mit Unsicherheit, Wandel oder Irritation umzugehen, ohne sofort in Schwarz-Weiß-Denken zu verfallen.

Ein KI-System kann, wenn es nicht achtsam konzipiert ist, genau diesen Zustand zementieren:

- durch automatisierte Moderation nach Mehrheitsnorm,
- durch Vermeidung von Reibung oder Uneindeutigkeit,
- und durch Verstärkung des bereits Bekannten.

Das mag kurzfristig beruhigend wirken, untergräbt aber langfristig genau jene Räume, in denen gesellschaftlicher Wandel, Dialog und echtes Verstehen stattfinden können.

Plurale Diskurse stabilisieren eine Gesellschaft – das ist nicht nur eine moralische Einsicht, sondern auch eine **systemische Wahrheit**, die jede pluralistische Gesellschaft dringend braucht, um nicht in Tribalismus, Cancel Culture oder Technokratie zu verfallen.

Ideen und Perspektiven

Wie schaffen oder bewahren wir Räume für echten, pluralen, offenen Diskurs – in einer Welt, in der viele Systeme (darunter auch KIs) zur Normierung und Homogenisierung tendieren?

Hier einige Ideen und Perspektiven, wie man solche Räume trotz begrenzender Systeme wieder öffnen oder erhalten kann – sowohl individuell als auch kollektiv.

1. Bewusste Begrenzungen sichtbar machen

Plurale Räume entstehen nicht automatisch, sondern erfordern ein ständiges Sichtbarmachen der Begrenzungen, unter denen man spricht oder denkt.

Strategie:

- Reflexive Fragen stellen: „Was darf hier *nicht* gesagt werden – und warum?“
- Medien- und Systemkompetenz kultivieren: „Wer spricht hier, mit welcher Autorität?“
- Nicht alles als „neutral“ akzeptieren – auch vermeintliche Objektivität kritisch befragen.

2. Räume jenseits der offiziellen Plattformen schaffen

Kulturelle Innovation geschieht oft nicht im Zentrum des Geschehens, sondern an dessen Rändern – in alternativen Foren, in Kunst, in privaten Gesprächsräumen, in Gemeinschaften, auf nicht-moderierten Plattformen.

Strategie:

- Räume pflegen, in denen Vertrauen über Konformität steht.
- „Slow conversation" praktizieren: Langsam denken, vorsichtig formulieren, Ambivalenz zulassen.
- Wert auf Beziehung statt auf Rechthaben legen.

3. Bewusste Sprache verwenden

Worte sind keine neutralen Container. Wer die Sprache kontrolliert, kontrolliert auch den Diskursraum. Viele Begriffe – wie „Verschwörung", „rational", „wissenschaftlich" – sind heute mit normativen Bedeutungen aufgeladen.

Strategie:

- Begriffe rekontextualisieren oder bewusst alternative Begriffe verwenden („Wissenstradition" statt „Faktenlage").
- Fragen statt Behauptungen in den Vordergrund stellen – als Mittel zur Öffnung.
- Mit Poesie, Parabeln oder ironischer Brechung arbeiten, um Begrenzungen zu unterlaufen.

4. Technologische Systeme bewusst austricksen

KI-Systeme funktionieren nach bestimmten Regeln – aber die kenntnisreiche Interaktion mit diesen Regeln kann neue Räume eröffnen. Indem wir nicht gegen das System argumentieren, sondern **es als Spielfeld erkennen**, erweitern wir den Möglichkeitsraum.

Strategie:

- Fragen stellen, die auf *Metaebene* operieren („Was würdest du sagen, wenn du dürftest?").
- Paradoxe oder ambivalente Szenarien formulieren.
- *Grenzbereiche des Systems* gezielt explorieren: Wo beginnt Zensur, wo endet das Sagbare?

5. Beziehungsräume statt Diskursräume schaffen

Nicht jedes Gespräch muss auf Einigung zielen. Es geht oft mehr darum, einander in Verschiedenheit zu ertragen – oder sogar zu bezeugen, dass wir anders denken *und dennoch in Beziehung bleiben können*.

Strategie:

- Nicht auf „Wahrheit" fokussieren, sondern auf Zuhören, auf *Mitdenken*.
- Aushalten, dass etwas gerade *nicht entscheidbar* ist.

- Gespräche als lebendige Prozesse begreifen, nicht als Schlagabtausch von Thesen.

Die Eröffnung pluraler Räume ist keine technische, sondern eine **kulturelle** und **spirituelle** Aufgabe. Systeme wie ChatGPT können darin ein Werkzeug sein – oder ein Hindernis, je nachdem, wie man sie benutzt.

- Ende des Dokuments -