// KI-Hub VON STRUKTURAFLOW

IT-SECURITY

# Superposition: Warum KI-Modelle besser werden – und schwerer zu kontrollieren

Eine MIT-Studie erklärt erstmals den Mechanismus hinter dem Blackbox-Problem: Was Geschäftsführer über Superposition wissen sollten, bevor sie KI produktiv einsetzen.

AUTOR
Natascha Reiner

VERÖFFENTLICHT
30. April 2026

ONLINE LESEN
https://wissen.strukturaflow.it.com/superposition-ki-blackbox-kontrollierbarkeit/

„Wir wissen nicht, wie unsere eigenen KI-Systeme funktionieren.„

Das ist kein Eingeständnis eines überforderten Anwenders. Es ist ein Zitat von Dario Amodei — CEO von Anthropic, einem der führenden KI-Unternehmen der Welt, das Modelle wie Claude entwickelt.

Wer KI produktiv in seinem Betrieb einsetzt — oder es plant — sollte bei diesem Satz kurz innehalten.

Denn wenn selbst die Hersteller nicht vollständig erklären können, warum ihre Modelle tun, was sie tun: Wie soll ein Unternehmen diese Systeme dann zuverlässig kontrollieren? Wie soll es Fehler erkennen, bevor sie teuer werden? Wie soll es gegenüber Kunden, Behörden oder Versicherungen nachweisen, dass die Technologie verlässlich arbeitet?

Eine neue MIT-Studie, ausgezeichnet als Best Paper Runner-Up der NeurIPS 2025 — der wichtigsten KI-Forschungskonferenz weltweit — liefert erstmals eine mechanistische Erklärung dafür, warum KI-Modelle so schwer zu verstehen sind. Und warum sich das nicht so schnell ändern wird.

# Was Superposition bedeutet

## Das Platzproblem

Stellen Sie sich vor, Sie müssten eine Bibliothek mit 50.000 Büchern in einem Regal mit 3.000 Fächern unterbringen.

Genau das ist das Grundproblem jedes Sprachmodells.

Ein LLM muss Zehntausende von Tokens — Wörter, Konzepte, Bedeutungen — in einem internen Rechenraum repräsentieren, der nur einige Tausend Dimensionen hat. In einem dreidimensionalen Raum passen ohne gegenseitige Störung genau drei Konzepte. In einem Raum mit 4.000 Dimensionen: 4.000 Konzepte.

Aber ein modernes Sprachmodell muss Hunderttausende abstrakte Konzepte gleichzeitig verarbeiten.

Das geht sich nicht aus. Zumindest nicht auf dem klassischen Weg.

## Die Lösung, die das Problem erst sichtbar macht

Sprachmodelle umgehen diese Beschränkung, indem sie viele Konzepte gleichzeitig in denselben Dimensionen speichern. Die zugehörigen Repräsentationsvektoren überlappen sich dabei leicht — sie teilen sich den Platz, ohne vollständig identisch zu sein.

Dieses Quetschen mehrerer Bedeutungen in zu wenig Raum nennt die Forschung **Superposition**.

Das ist kein Designfehler. Es ist eine emergente Eigenschaft — etwas, das Modelle im Training selbst entwickeln, weil es funktioniert.

Und es funktioniert tatsächlich: Die MIT-Studie zeigt, dass echte Sprachmodelle ausnahmslos im Modus **starker Superposition** arbeiten. Sie quetschen nicht nur die häufigsten Konzepte rein — sie quetschen alles rein, inklusive seltener und abstrakter Bedeutungen. Alle gleichzeitig, in überlappenden Vektoren.

# Warum das Modelle besser macht – je größer, desto besser

## Das Geheimnis hinter den Skalierungsgesetzen

Eine der stabilsten Beobachtungen der KI-Forschung lautet: Verdoppelt man die Modellgröße, sinkt der Vorhersagefehler nach einem vorhersagbaren Muster. Mehr Parameter, bessere Ergebnisse. Mehr Trainingsdaten, bessere Ergebnisse. Mehr Rechenleistung, bessere Ergebnisse.

Diese sogenannten **Neural Scaling Laws** treiben seit Jahren den Bau immer größerer Modelle an — und damit Milliarden an Investitionen. Warum sie gelten, war bisher empirisch beobachtbar, aber theoretisch nicht vollständig erklärt.

Die MIT-Studie schließt diese Lücke.

## Was das MIT herausgefunden hat

Der Mechanismus ist direkt aus der Geometrie der Superposition ableitbar: Je mehr Dimensionen ein Modell hat, desto weniger überlappen sich die Vektoren für einzelne Konzepte. Weniger Überlappung bedeutet weniger Rauschen. Weniger Rauschen bedeutet präzisere Ausgaben.

Die Formel ist schlicht: Verdoppelt man die Modellbreite, halbiert sich der Fehler durch Überlappungen — unabhängig davon, wie die Trainingsdaten verteilt sind. Das ist der Grund, warum Skalierung so verlässlich funktioniert.

Die Studie wurde nicht nur auf Spielzeugmodellen bestätigt. Das Team hat reale Open-Source-Sprachmodelle untersucht — OPT, GPT-2, Qwen, Pythia — mit Größen von 100 Millionen bis 70 Milliarden Parametern. Das Ergebnis war eindeutig: Alle arbeiten im Regime der starken Superposition. Die Skalierungsgesetze gelten, weil die Geometrie es so erzwingt.

Und wann endet Skalierung? Laut den Forschenden dann, wenn ein Modell breit genug ist, um jedes Konzept ohne Überlappung abzubilden. An dieser Grenze verliert weiteres Wachstum seinen Effekt. Für natürliche Sprache ist diese Grenze noch nicht erreicht.

# Die andere Seite der Gleichung

## Mehr Leistung, weniger Einblick

Superposition macht Modelle leistungsfähiger. Sie macht sie gleichzeitig undurchsichtiger.

Das ist kein Zufall. Es ist dieselbe Medaille.

Wenn Konzepte in überlappenden Vektoren gespeichert sind, lässt sich von außen nicht mehr sauber trennen, welche Konzepte an einer bestimmten Ausgabe beteiligt waren. Das Modell trifft eine Entscheidung — aber die interne Repräsentation, die zu dieser Entscheidung geführt hat, ist ein Gewirr aus Tausenden überlappender Signale.

Je mehr Superposition, desto besser die Leistung. Je mehr Superposition, desto schwerer die Interpretierbarkeit.

Das ist der strukturelle Kern des Blackbox-Problems — und er ist nicht durch ein Update lösbar.

## Der Versuch, die Blackbox zu öffnen

Die Forschungsdisziplin, die sich mit genau diesem Problem beschäftigt, heißt **Mechanistic Interpretability**. Ziel: Neuronale Netze von innen verstehen — nicht nur beobachten, was sie ausgeben, sondern nachvollziehen, wie sie intern rechnen.

MIT Technology Review hat Mechanistic Interpretability im Februar 2026 als eine der **10 wichtigsten Breakthrough Technologies des Jahres** ausgezeichnet. Anthropic hat Werkzeuge entwickelt, die wie ein Mikroskop ins Modell schauen — und einzelne Konzepte und Pfade sichtbar machen können. OpenAI und Google DeepMind nutzen ähnliche Techniken, um unerwartetes Verhalten ihrer Modelle zu untersuchen.

Der Fortschritt ist real. Aber er ist langsam.

Ein aktueller Statusbericht aus der Forschungs-Community fasst die Lage so zusammen: Grundlegende Konzepte wie „Feature“ haben noch keine rigorosen Definitionen. Viele Fragen zur Interpretierbarkeit sind rechnerisch nicht lösbar. Und praktische Methoden schneiden bei sicherheitsrelevanten Aufgaben noch schlechter ab als einfache Baselines.

Anthropic hat als Ziel formuliert, bis 2027 die meisten Probleme in KI-Modellen zuverlässig erkennen zu können. Das ist ein ambitioniertes Versprechen — und ein ehrliches Eingeständnis, dass es heute noch nicht gilt.

# Was das für Unternehmen konkret bedeutet

## Vertrauen ohne Verstehen

Wenn ein Unternehmen einen KI-Agenten auf sein CRM, seine Kundenkommunikation oder seine interne Wissensdatenbank loslässt, vertraut es einem System, dessen interne Logik selbst die Hersteller nicht vollständig entschlüsseln können.

Das ist keine Panikmache. Es ist die nüchterne Beschreibung des aktuellen Stands.

Und es verändert die Fragen, die Geschäftsführungen stellen sollten. Nicht nur: „Funktioniert das? „ Sondern: „Wie merken wir, wenn es aufhört zu funktionieren?“ Und: „Wie erklären wir das, wenn jemand fragt?„

## Die Kontrollierbarkeits-Frage

Der EU AI Act verlangt von Unternehmen, die KI in bestimmten Anwendungsbereichen einsetzen, Nachvollziehbarkeit und Dokumentation. Das ist eine politische Antwort auf ein technisches Problem.

Das technische Problem ist Superposition.

Wer nicht erklären kann, warum ein Modell eine bestimmte Ausgabe produziert hat, kann diese Ausgabe auch nicht zuverlässig kontrollieren. Das ist keine theoretische Schwäche. Es hat praktische Konsequenzen: bei Fehlentscheidungen, bei Haftungsfragen, bei Audits.

Die Anforderung an Unternehmen ist deshalb nicht, die Interna von Sprachmodellen zu verstehen. Das kann heute niemand vollständig.

Die Anforderung ist, das System so zu gestalten, dass mangelnde Interpretierbarkeit nicht unkontrolliert zum Risiko wird.

## Was das in der Praxis bedeutet

Konkret heißt das für jeden Betrieb, der KI produktiv einsetzt oder plant:

**Outputs überprüfen, nicht nur Inputs definieren.** Ein Modell, das gute Ergebnisse in einer Testumgebung liefert, tut das nicht zwingend dauerhaft in variablen Realbedingungen. Wer keine Struktur hat, um Outputs regelmäßig zu bewerten, hat kein Kontrollsystem.

**Irreversible Aktionen absichern.** Wo ein KI-Agent Entscheidungen mit echten Konsequenzen trifft — Kundenkommunikation, Datenbankzugriffe, Finanztransaktionen — braucht es menschliche Kontrollpunkte. Nicht weil das Modell bösartig ist. Sondern weil seine interne Logik nicht vollständig nachvollziehbar ist.

**Dokumentation als Schutz.** Was hat der Agent getan, wann, auf welcher Grundlage? Logs sind kein bürokratischer Aufwand. Sie sind im Streitfall das einzige, was eine Entscheidung rekonstruierbar macht.

**Den richtigen Einsatzbereich wählen.** Superposition und Blackbox-Problematik bedeuten nicht, dass KI nicht eingesetzt werden soll. Sie bedeuten, dass der Einsatz in klar abgegrenzten, überschaubaren Prozessen mit niedrigem Fehlerpotenzial anders zu bewerten ist als in kritischen Entscheidungsprozessen mit hoher Fehlerkonsequenz.

## Was sich gerade verändert

Die Forschung zu Mechanistic Interpretability wächst schnell. Vor fünf Jahren arbeiteten eine Handvoll Forschende daran. Heute sind es Hunderte, verteilt auf die führenden KI-Labore weltweit.

Das ist relevant — nicht weil das Blackbox-Problem morgen gelöst sein wird, sondern weil die Werkzeuge zur Kontrolle und Prüfung von KI-Systemen in den nächsten Jahren deutlich besser werden. Unternehmen, die heute eine saubere Grundarchitektur aufbauen, werden von diesen Werkzeugen profitieren können, sobald sie verfügbar sind.

Wer heute hingegen KI-Agenten ohne Monitoring, ohne Dokumentation und ohne klare Abgrenzung ihrer Handlungsfreiheit einsetzt, hat nicht nur ein aktuelles Risiko. Er baut eine Struktur auf, die sich nachträglich nur schwer korrigieren lässt.

Das Blackbox-Problem ist kein Grund, KI nicht einzusetzen. Es ist ein Grund, es strukturiert zu tun.

## Wo stehen Sie mit Ihrer KI-Architektur?

Wenn Sie einschätzen möchten, wie Ihre bestehenden oder geplanten KI-Systeme in Bezug auf Kontrolle, Dokumentation und Risikobegrenzung aufgestellt sind — und wo konkreter Handlungsbedarf besteht:

Jetzt kostenlos Ihr Potenzial analysieren.

NÄCHSTER SCHRITT

### Mehr praktische KI-Anleitungen für KMU

Dieser Artikel ist Teil des KI-Hubs von Strukturaflow — einer deutschsprachigen Plattform für den praktischen KI-Einsatz in kleinen und mittleren Unternehmen.

https://wissen.strukturaflow.it.com