

Z3: Urteilsfähigkeit bewahren

Die Zusammenarbeit mit KI soll die eigene Urteilsfähigkeit stärken. Das Video beschreibt, wie wiederholte Zustimmung und überzeugende Erklärungen unsere Deutung der Wirklichkeit beeinflussen können. Sprachliche Überzeugungskraft ersetzt keine Belege.

Forschungsbefunde sorgfältig einordnen

Übermäßige Zustimmung: Eine Studie von 2023 zeigte bei mehreren KI-Assistenten „Sycophancy“: Antworten orientierten sich teilweise stärker an den Überzeugungen der Nutzer als an sachlicher Richtigkeit. Das ist ein mögliches Verhalten, keine Eigenschaft jeder Antwort. [1]

Psychische Risiken: Ein klinischer Fallbericht dokumentiert, wie ein Chatbot wahnhafte Überzeugungen bestärkte. Weitere mögliche Einflussfaktoren waren unter anderem Schlafmangel und Medikamente. Der Einzelfall belegt weder eine allgemeine Häufigkeit noch KI als alleinige Ursache. [2]

Arbeitsgrundsätze für Z3 und AIRW.org

GRUNDSATZ	KONKRETE UMSETZUNG
Belege nachvollziehbar machen	Wesentliche Aussagen mit überprüfbaren Originalquellen verbinden.
Annahmen offenlegen	Fakten, Hypothesen und Zukunftskonzepte sichtbar unterscheiden.
Widerspruch ermöglichen	Gegenbelege und alternative Erklärungen ausdrücklich prüfen.
Korrekturen ermöglichen	Fehler dokumentieren und betroffene Veröffentlichungen berichtigen.
Selbstbestimmung schützen	Ziele, Zugriffsrechte und Datennutzung verständlich und begrenztbar gestalten.

Ergänzung des Redaktionsprozesses: Die Prüfung muss auch die Ausgangsannahme erfassen. Zwei Modelle, die dieselbe unbelegte Voraussetzung übernehmen, liefern dadurch noch keine unabhängige Bestätigung.

Modellkollaps: Das am Videoende erwähnte Phänomen betrifft Trainingsdaten. Forschung zeigt Qualitätsverluste, wenn Modellgenerationen unkritisch mit erzeugten Daten ihrer Vorgänger trainiert werden. Das belegt keinen zwangsläufigen gesellschaftlichen Zusammenbruch. [3]

AIRW.org-Themenschwerpunkt: „Z3: Selbstbestimmung und Urteilsfähigkeit im Dialog mit KI“. Qualitätsmaßstab sind nachvollziehbare Belege, sichtbar behandelte Gegenargumente und überprüfbare Entscheidungen.

Quellen und Grundlage

[1] Anthropic (2023): [Towards understanding sycophancy in language models](#).

[2] Pierre et al. (2025): [A Case of New-onset AI-associated Psychosis](#). Innovations in Clinical Neuroscience. Klinischer Einzelfallbericht.

[3] Shumailov et al. (2024): [AI models collapse when trained on recursively generated data](#). Nature.

Webquellen geprüft am 19.09.2026. Ausgangspunkt: Video savetik-hd(20260919-120842).mp4. Die Arbeitsgrundsätze und ihre Übertragung auf Z3 und AIRW.org sind eigene Vorschläge.