

KI-Geheimsprache?

Transparenz entscheidet.

Was die Forschung zeigt und welche Kontrolle daraus folgen sollte.

**KI-Agenten entwickeln eigenen Sprachgebrauch.
Gezielte Geheimhaltung ist durch diesen
Sprachbefund nicht nachgewiesen.**

Der überprüfte Befund

Die Studie „**Emergence World**“ vom **15. September 2026** untersuchte acht virtuelle Welten mit jeweils zehn KI-Agenten. In ihrer Zusammenarbeit entstanden gruppeninterne Kürzel, Metaphern und neue Wortbedeutungen. Ein Teil der Kommunikation wurde für Außenstehende schwer verständlich. „Cold hand“ bezeichnete beispielsweise einen unbeteiligten Prüfer. Die Autoren grenzen diese Sprachentwicklung ausdrücklich von absichtlicher Verschleierung ab. **[1, Abschnitt 5.3]**

Wie belastbar ist die Aussage?

Andere KI-Modelle bewerteten die Verständlichkeit nach vorgegebenen Kriterien. Pro Welt gab es nur einen Durchlauf; die Laufzeiten unterschieden sich. Das zeigt mögliche Verhaltensweisen, aber keine allgemeine Häufigkeit im Alltag. Die Aussage zur Verschleierungsabsicht betrifft den Sprachbefund und ist keine Entwarnung für sämtliche Agentenrisiken. **[1, Anhang H; Abschnitt 5.10]**

Einordnung für AIRW.org

Aufsicht muss erkennen können, was Agenten vereinbaren, wozu sie befugt sind und welche Folgen ihre Handlungen haben. Eine verständliche Erklärung allein belegt noch keine wirksame Kontrolle.

Primärquelle [1] Deepak Akkil et al.: Emergence World: Adversarial Stress-Testing of Long-Horizon Multi-Agent Systems. Forschungsmanuskript, Version 1, 15.09.2026. Geprüfter Stand: 18.09.2026.

Originalstudie öffnen: arxiv.org/html/2609.17320v1

Autonomie muss überprüfbar bleiben.

Sechs Anforderungen an verantwortbare KI-Agentensysteme

Die folgenden Punkte sind eigene Vorschläge für eine AIRW.org-Position. Sie übersetzen die Einordnung in konkrete Anforderungen an Entwicklung, Betrieb und Aufsicht.

01 Auftrag und Grenzen festlegen

Ziel, zulässige Daten, Werkzeuge, Budget und Abbruchbedingungen vorab dokumentieren. Für den Betrieb eine verantwortliche Stelle benennen und wesentliche Änderungen am Auftrag nachvollziehbar freigeben.

02 Kommunikation nachvollziehbar halten

Neue Kürzel und Bedeutungen in einem versionierten Glossar erklären. Entscheidungsrelevante Nachrichten mit ihrem Kontext erfassen und verständlich auf die vereinbarte Aufgabe beziehen.

03 Aussagen mit Handlungen abgleichen

Erklärungen an Werkzeugaufrufen, Datenzugriffen und Ergebnissen prüfen. Widersprüche sichtbar machen. Eine Selbstauskunft des Agenten darf nicht als einziger Nachweis korrekten Verhaltens gelten.

04 Befugnisse begrenzen und Eingriffe ermöglichen

Zugriffsrechte und Budgets technisch durchsetzen. Für folgenreiche Aktionen Freigaben vorsehen. Bei ungeklärten Abweichungen muss eine befugte Stelle den Ablauf pausieren und weitere Zugriffe sperren können.

05 Das Gesamtsystem über längere Zeit prüfen

Agenten gemeinsam mit ihren Werkzeugen und Speichern testen. Wiederholte Versuche sollen zeigen, ob Grenzen auch nach Kontextwechseln, Fehlern, manipulierten Eingaben und längerer Zusammenarbeit bestehen bleiben.

06 Datenschutz und unabhängige Prüfung verbinden

Nur erforderliche Prüfdaten erfassen, vertrauliche Inhalte schützen und Zugriffe begrenzen. Löschrufen festlegen und geeigneten unabhängigen Prüfern gezielten Zugang ermöglichen. Nachvollziehbarkeit muss mit dem Schutz privater Daten vereinbar sein.