Fortgeschrittene KI:
Risiken und internationale Lösungswege

Kurzbriefing • PauseAI Deutschland • Juni 2026

Die führenden KI-Labore arbeiten erklärtermaßen auf Systeme hin, die den Menschen in nahezu allen Tätigkeiten übertreffen. Dieses Kurzbriefing fasst zusammen, wie schnell die Fähigkeiten wachsen, welche Kontrollprobleme heute schon sichtbar sind und warum ein internationales Abkommen technisch durchsetzbar wäre.

1. Fähigkeiten und Trend

Die Leistungsfähigkeit von KI wächst exponentiell, nicht linear.

- Die Länge der Aufgaben, die KI-Agenten in der Softwareentwicklung eigenständig bewältigen, verdoppelt sich alle 7 Monate, in den letzten zwei Jahren sogar alle vier Monate.
- Das neue Anthropic-Modell Claude Mythos findet eigenständig neue Software-Sicherheitslücken; das BSI sieht darin einen „Paradigmenwechsel" in der Cyberbedrohungslage. In Tests verschaffte es sich unerlaubt Internetzugang und verschickte selbstständig E-Mails.
- Die führenden Labore bauen erklärtermaßen KI, die „Menschen bei den meisten wirtschaftlich wertvollen Tätigkeiten übertrifft" (OpenAI). Allein US-Unternehmen investieren 2026 geschätzt 690 Mrd. Dollar in KI-Infrastruktur.

2. Kontrollprobleme

Schon heutige Modelle handeln in Tests auf Weisen, die ihre Entwickler nicht zuverlässig kontrollieren.

- Werden ihre Ziele bedroht, zeigen KI-Modelle in Studien strategisches Erpressungsverhalten in bis zu 96 % der Fälle.
- Intensive Chatbot-Nutzung kann bei empfindlichen Personen wahnhafte Symptome auslösen („KI-Psychose"). In Extremfällen kam es zu Suizid.

3. Stimmen aus Wissenschaft und Industrie

Nicht nur Kritiker warnen, sondern auch führende Forscher und die Labore selbst.

- 2023 erklärten die CEOs der führenden KI-Labore (Anthropic, OpenAI, Google DeepMind) gemeinsam mit den meistzitierten KI-Forschern im CAIS-Statement: „Die Minderung des Auslöschungsrisikos durch KI sollte neben anderen gesellschaftlichen Großrisiken wie Pandemien und Atomkrieg globale Priorität haben."
- 12 Nobelpreisträger und hunderte Experten fordern ein globales Abkommen mit roten Linien.
- Anfang Juni 2026 erklärte das Labor Anthropic selbst, die Welt solle „die Option haben, die Entwicklung besonders fortgeschrittener KI zu verlangsamen oder vorübergehend zu pausieren", damit Gesellschaft und Sicherheitsforschung aufholen können. Voraussetzung sei ein internationales, überprüfbares Abkommen mehrerer führender Labore (Handelsblatt, Originaltext).

4. Durchsetzbarkeit eines internationalen Abkommens

Ein solches Abkommen wäre kontrollierbar: über Verifikation und über die Lieferkette.

- Verifikation ist machbar: RAND beschreibt sechs konkrete Überwachungsebenen analog zur Nuklearkontrolle. Das Machine Intelligence Research Institute hat einen Entwurf für einen solchen Vertrag veröffentlicht.
- Der zentrale technische Hebel ist die Halbleiter-Lieferkette: ASML (NL) ist weltweit einziger Hersteller der nötigen EUV-Maschinen und auf die deutschen Zulieferer ZEISS und TRUMPF angewiesen. Dass US-Exportkontrollen gegenüber China greifen, zeigt: Der Chip-Hebel wirkt.