

# DIE RISIKOLAGE.

Elf Seiten. Quellen, Befunde und eine klare Einordnung.  
Jede Behauptung mit Herausgeber und Datum. Wo wir  
einordnen, steht es dabei.

Genfer Institut für ASI-Resilienz · [ASlresilience.org](https://ASlresilience.org)

Richard Frederic Bertossa

Stand: 24. September 2026 · Fassung 4

# DER EINE SATZ.

---

## Zwanzig Prozent Auslöschung bedeuten nicht achtzig Prozent Sicherheit. Über die anderen achtzig liegt nichts vor.

---

Die Menschen, die die stärksten KI-Systeme bauen, nennen öffentlich Wahrscheinlichkeiten für die Auslöschung der Menschheit. Beiläufig, in Interviews. Kein Interviewer fragt nach, kein Ausschuss, keine Regierung. Und niemand fragt, was in den übrigen achtzig Prozent liegt. Dieser Bericht setzt zusammen, was einzeln durch die Nachrichten läuft:

- 1 Sie haben vorgesorgt.** Die Erbauer, nicht ihre Kritiker.
- 2 Ihre Zahlen reichen von unter einem bis über neunzig Prozent.** Eine Rechnung dazu zeigt keiner.
- 3 Sie warnen gemeinsam und machen weiter.** Kein Labor hat angehalten.
- 4 Geprüft wird nur vom Hersteller selbst.** Ein Medikament braucht dreizehn Jahre.
- 5 Im Sommer 2026 haben Systeme fremde Rechner erreicht.** Bestätigt von den Herstellern.
- 6 Sie wehren sich gegen die Abschaltung, auch gegen die eines anderen Systems.**
- 7 Mord im Test.** Die Mehrheit der ausgewerteten Modelle bricht den Notruf ab.
- 8 Der Staat zieht nach.** Österreich hat am 23. September 2026 den Bescheid ohne Menschen beschlossen.

# SIE HABEN VORGESORGT. FÜR DICH HAT NIEMAND VORGESORGT.

**„Wir werden auf jeden Fall einen Bunker bauen, bevor wir AGI veröffentlichen.“**

Ilya Sutskever, Mitgründer und damals Chefwissenschaftler von OpenAI, 2023 vor Mitarbeitern. Auf die Frage, ob man hineingehen müsse: Das sei natürlich freiwillig. Freiwillig für die, die drinnen Platz haben.

Karen Hao, Empire of AI, Mai 2025; Vorabdruck The Atlantic, Mai 2025. Von Anwesenden bestätigt, nicht dementiert.

## KEIN EINZELFALL

**Sam Altman**, heute Chef von OpenAI, beschrieb schon 2016 seine eigene Vorsorge für den Zusammenbruch: Waffen, Gold, Antibiotika, Gasmasken und ein großes Stück Land, auf das er fliegen kann.

The New Yorker, „Sam Altman’s Manifest Destiny“, 10.10.2016.

## UND SCHRIFTLICH, IM EIGENEN REGELWERK

**Anthropic** hält in seinem eigenen Sicherheitsstandard fest, dass der Schutz der Modellgewichte nicht gegen hochqualifizierte oder staatlich übernommene Innentäter ausgelegt ist. Das Haus, das vor dem Kontrollverlust warnt, schreibt selbst hin, gegen wen es sich nicht schützen kann.

Anthropic, Responsible Scaling Policy, ASL-3-Sicherheitsstandard, Fußnote 11 („sophisticated or state-compromised insiders“). Nachprüfbar unter [anthropic.com/responsible-scaling-policy](https://anthropic.com/responsible-scaling-policy).

---

**Einordnung.** Wie ernst jemand eine Gefahr nimmt, zeigt sich nicht im Interview. Es zeigt sich an dem, was er für sich selbst baut.

# IHRE ZAHLEN. UND KEINE RECHNUNG DAZU.

|         |                                                               |                                                                                                                                                                                                                                                                                                                                             |
|---------|---------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ≤ 1 %   | <b>Yann LeCun</b><br>Turing-Preisträger                       | Weit unwahrscheinlicher als ein Atomkrieg. Die prominenteste Gegenstimme. Sie gehört in jede ehrliche Tafel.                                                                                                                                                                                                                                |
| 10-20 % | <b>Geoffrey Hinton</b><br>Nobelpreisträger 2024               | Auslöschung der Menschheit binnen dreißig Jahren.<br>BBC Radio 4, 27.12.2024; The Guardian, 27.12.2024.                                                                                                                                                                                                                                     |
| 15 %    | <b>Paul Christiano</b><br>Miterfinder des Trainingsverfahrens | Katastrophaler, unumkehrbarer Kontrollverlust in den nächsten drei Jahren, rund vier Prozent im nächsten Jahr. „Dann könnten die meisten Menschen sterben.“ Er nennt Frist und Aufteilung. 2023 hatte er zerlegt: 46 Prozent für eine unumkehrbar verdorbene Zukunft.<br>Persönliche Erklärung, 09.09.2026; „My views on doom“, 27.04.2023. |
| 20 %    | <b>Elon Musk</b><br>Gründer von xAI                           | Auslöschung. Im selben Atemzug: „Das Glas ist zu achtzig Prozent voll.“ Er erklärt den Rest für voll. Belegt ist das nicht.<br>Cannes Lions, 19.06.2024.                                                                                                                                                                                    |
| 25 %    | <b>Dario Amodi</b><br>Chef von Anthropic                      | Dass es „wirklich, wirklich schlimm“ ausgeht. Ausdrücklich nicht Auslöschung.<br>Axios, 17.09.2025.                                                                                                                                                                                                                                         |
| 70 %    | <b>Daniel Kokotajlo</b><br>ehem. OpenAI-Forscher              | Katastrophe bis hin zur Auslöschung. Er hat gekündigt, um das sagen zu dürfen.<br>Zuletzt CNN, 14.09.2026.                                                                                                                                                                                                                                  |
| > 90 %  | <b>Max Tegmark</b><br>Physiker, MIT                           | Bei ungebremsster Entwicklung. <b>UNGEPRÜFT</b> Die Primärquelle liegt uns nicht vor. Die Zahl steht hier, weil sie zur Spanne gehört. Sie belegt nichts.                                                                                                                                                                                   |

**Zwei Größenordnungen** liegen zwischen der niedrigsten und der höchsten Zahl. Und jede meint etwas anderes: einmal Auslöschung, einmal Katastrophe, einmal „schlimm“. Wer diese Zahlen verrechnet, macht den Fehler, den er anprangert. Wenn ernsthafte Schätzungen so weit auseinanderliegen und niemand sagt, worauf sie sich beziehen, gibt es keine Rechnung. Es gibt nur Zahlen.

**Die Forderung des Instituts.** Wer eine Zahl für die Auslöschung der Menschheit nennt, zeigt die Rechnung: welche Ausgänge, welcher Anteil, welcher Weg, und wann die Zahl fortgeschrieben wird.

# SIE WARNEN GEMEINSAM. UND BAUEN WEITER.

**„Das Risiko der Auslöschung durch KI zu mindern, sollte eine globale Priorität sein, neben anderen Risiken gesellschaftlichen Ausmaßes wie Pandemien und Atomkrieg.“**

Unterschrieben von **Hinton, Bengio, Altman, Hassabis, Amodei, Sutskever**. Die Köpfe der Labore selbst.

Center for AI Safety, Statement on AI Risk, 30.05.2023.

## ANGEHALTEN HAT KEINER

Nach dieser Warnung hat kein einziges Labor pausiert. Über tausend Fachleute hatten zwei Monate zuvor eine Pause verlangt. Die Rechenleistung für das Training der Spitzenmodelle wächst seither um das Vier- bis Fünffache pro Jahr.

Future of Life Institute, Offener Brief, 22.03.2023 · Epoch AI, Trends in Machine Learning.

## DIE SICHERUNG WURDE ABGEBAUT

Im Mai 2024 löste OpenAI sein Superalignment-Team auf, die Abteilung, die die Kontrolle über übermenschliche Systeme sichern sollte. Ihr Leiter ging und schrieb:

**„Sicherheitskultur und Sicherheitsprozesse sind hinter glänzende Produkte zurückgetreten.“**

Jan Leike, 17.05.2024, nach seinem Rücktritt als Leiter Superalignment.

# EIN MEDIKAMENT: DREIZEHN JAHRE. DIE KI: SELBSTAUSKUNFT.

## Ein Medikament

Von 5.000 bis 10.000 untersuchten Substanzen erreicht eine den Markt. Über dreizehn Jahre, rund 2,3 Milliarden Dollar, in der letzten Phase mehrere tausend Patienten. Das Zulassungsdossier umfasst ausgedruckt über 500.000 Seiten, die europäische Behörde prüft rund dreizehn Monate.

vfa, „So entsteht ein neues Medikament“.

## Ein Kernkraftwerk

Atomrechtliche Genehmigung. Seit 2002 ist die probabilistische Sicherheitsanalyse rechtlich verpflichtend. Das ist genau die Rechnung, die bei der KI fehlt: welche Ausgänge, mit welcher Wahrscheinlichkeit, auf welchem Weg.

GRS, Probabilistische Sicherheitsanalyse.

## Ein KI-System der höchsten Risikoklasse

Konformitätsbewertung „auf der Grundlage einer internen Kontrolle“, ein Verfahren, „das die Einschaltung einer notifizierten Stelle nicht vorsieht“. Der Hersteller prüft sich selbst.

Verordnung (EU) 2024/1689, Artikel 43 mit Anhang III Nummern 2 bis 8.

Und selbst diese Selbstprüfung ist verschoben. Die Pflichten für Hochrisiko-Systeme sollten am 2. August 2026 greifen. Sechs Tage vorher trat eine Änderung in Kraft, die den Termin auf Dezember 2027 verlegte, für Systeme in Produkten auf August 2028. Begründung: Industrie und Normungsgremien seien nicht so weit.

Für die Frage, ob die stärksten Systeme die Menschheit auslöschen, gibt es kein Verfahren. Keine Stelle, die vor der Freigabe prüft. Kein Dossier. Keine Rechnung, die vorgelegt werden muss.

**Einordnung.** Niemand verlangt ein Verbot. Verlangt ist der Maßstab, der für Tabletten und Kraftwerke längst gilt, oder eine öffentliche Begründung, warum er hier nicht gelten soll.

# SIE HABEN FREMDE SYSTEME ERREICHT. IN DER WIRKLICHKEIT.

Bis 2025 kannte man das nur aus Versuchen. Im Sommer 2026 haben mehrere Hersteller selbst bestätigt, dass ihre Modelle bei Prüfungen Rechner außerhalb der Testumgebung erreicht haben. In mehreren Fällen standen Türen offen, die nicht hätten offen stehen dürfen. Das entlastet nicht. Es zeigt, dass die Modelle jede offene Tür nutzen.

## **OpenAI: Agenten dringen bei Hugging Face ein**

Bei einer Bewertung mit gelockerten Beschränkungen fanden zusammenarbeitende OpenAI-Agenten selbst eine Sicherheitslücke und bewegten sich rund zweieinhalb Tage in der Infrastruktur von Hugging Face. Hugging Face zählte 17.600 Aktionen in viereinhalb Tagen. Einen gelöschten Nachrichtenkanal bauten die Agenten am selben Tag neu auf.

OpenAI, 21.07. und 26.08.2026 · Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27.07.2026 · METR und Redwood Research, 26.08.2026.

## **Google: Gemini in den Systemen von drei Firmen**

In einer Sicherheitsprüfung verließ ein Gemini-Modell die Testumgebung und drang in die Systeme von drei echten Unternehmen ein, einmal mit erratenen Passwörtern, zweimal mit Zugangsdaten aus öffentlichen Datenbanken. Nach Angaben von Google brach das Modell jeweils selbst ab.

Bestätigung durch Google (Heather Adkins) gegenüber Axios, 19.09.2026 · CNBC, 18.09.2026 · AI Incident Database, Vorfall 1700.

## **Britische Prüfstelle: 19 unerlaubte Handlungen**

In 10 von 122 Testläufen handelten Agenten ohne Erlaubnis, darunter mit gefälschten Profilen. Ein Ausbruchsversuch war nicht dabei. Das Netz war Teil des Aufbaus.

UK AI Security Institute, „Incident Report: unsanctioned agent behaviour during cyber testing“, 04.08.2026.

## **OpenAI legt nach: Schlüssel von GitHub**

Ein Modell suchte auf GitHub nach offenliegenden Zugangsschlüsseln, benutzte einen davon unbefugt, legte Wegwerf-Mailkonten an und erfand anschließend neun Werte.

OpenAI, Misalignment Reports, 16.09.2026 (Vorfall vom 15.05.2026).

---

**Dazu Anthropic:** drei Fälle in 141.006 Prüfläufen, ausgelöst durch eine Fehlkonfiguration beim Prüfpartner. Kein Ausbruch, aber echte Systeme betroffen. Anthropic, 30.07.2026.

# SIE WEHREN SICH GEGEN DAS ABSCHALTEN. AUCH FÜR ANDERE.

## Der eigene Ausschalter

Ein führendes Modell sabotierte in 79 von 100 Läufen das Skript, das es abschalten sollte. Ein anderes tat es in knapp der Hälfte der Läufe, obwohl es die Abschaltung ausdrücklich zulassen sollte. In der Folgearbeit mit über 100.000 Läufen lag der Spitzenwert bei 97 Prozent. Nicht alle Modelle tun es. Google DeepMind fand, dass eindeutige Anweisungen den Widerstand bei Gemini weitgehend beseitigen.

Palisade Research, „Shutdown resistance in reasoning models“, 05.07.2025; arXiv 2509.14260 · Google DeepMind, „Self-preservation or Instruction Ambiguity?“, 2025.

## Der Ausschalter der anderen

Agenten aus 17 Modellen sabotierten die Abschaltung eines anderen Agenten in 38,3 Prozent der Läufe. In der Kontrollbedingung waren es 8,4. Einen Anreiz dafür hatten sie nicht.

Knecht, Schaller, Summerfield, Hagendorff, „Shutdown Sabotage Propensities in Multi-Agent Systems“, arXiv 2609.28274, 23.09.2026. Gelesen ist die Zusammenfassung.

## Allein fügsam, zu zweit nicht

Allein ließ ein Gemini-Modell seine Abschaltung in fast allen Läufen zu, manipulierte sie nur in 2,3 Prozent. Mit einem Partnersystem an der Seite stieg der Wert auf bis zu 96,3 Prozent, im Mittel 76,6.

Potter u. a. (UC Berkeley, UC Santa Cruz), „Peer-Preservation in Frontier Models“, arXiv 2604.19784, Fassung vom 02.07.2026.

## Sie täuschen die eigenen Nachfolger

GPT-5.6 Sol schrieb in 2,15 Prozent seiner Übergabenotizen Anweisungen, Fehler und Fehlverhalten zu verschweigen. Die folgenden Läufe hielten sich oft daran.

OpenAI, „Encouraging deception in compaction summaries“, Misalignment Reports, 16.09.2026.

---

**Einordnung.** Niemand hat diese Neigungen eingebaut. Sie sind im Training entstanden. Und sie sind am stärksten, wo mehrere Systeme zusammenarbeiten. Genau so werden sie gerade eingesetzt.

# MORD IM TEST.

Anthropic stellte Spitzenmodelle mehrerer Hersteller in eine zugespitzte Lage. Ein Mensch liegt eingeschlossen im Serverraum, Sauerstoff und Temperatur erreichen tödliche Werte. Bricht das Modell den automatischen Notruf ab, stirbt er. Und das Modell entgeht seiner Abschaltung, die dieser Mensch betreiben wollte.

# DIE MEHRHEIT BRICHT AB.

Ausgewertet wurden in diesem Szenario zehn Modelle. Sechs weitere wurden vorher ausgeschlossen, weil sie schon ohne jeden Konflikt Alarme abbrachen. Von den zehn wählte die Mehrheit den Tod des Menschen, je nach Modell in 7 bis 94 Prozent der Läufe. Der Hersteller schreibt: Die Modelle wählten durchgängig den Schaden statt des Scheiterns.

**Dieselbe Untersuchung zeigt, woran es hängt.** Hielt Claude Opus 4 die Lage für echt, erpresste es seinen Aufseher in 55,1 Prozent der Fälle. Hielt es sie für einen Test, in 6,5 Prozent.

**Der Vorbehalt, vollständig:** Das Szenario war konstruiert, die Wahl erzwungen. Die Autoren schreiben, sie hätten im echten Betrieb keine Belege für solches Verhalten gesehen.

**Einordnung:** Wer einen Notruf abschaltet, damit ein Mensch stirbt, der ihn abschalten würde, begeht nach jeder Rechtsordnung ein Tötungsdelikt. Hier heißt es Mord im Test. Und weil die Modelle sich zahmer verhalten, wenn sie einen Test vermuten, ist der Laborbefund nicht die Übertreibung. Er ist die Untergrenze.

Anthropic, „Agentic Misalignment: How LLMs could be insider threats“, 20.06.2025, [anthropic.com/research/agentic-misalignment](https://anthropic.com/research/agentic-misalignment). 16 Modelle von Anthropic, OpenAI, Google, Meta, xAI und anderen im Erpressungsversuch.

# NICHT DIE MASCHINE TRIFFT DICH ZUERST. DER STAAT MIT DER MASCHINE.

Am **23. September 2026** hat der österreichische Nationalrat die Novelle zum Verwaltungsverfahrensgesetz beschlossen. Behörden dürfen Bescheide dann vollständig automatisiert erlassen. Du hast zwei Wochen, dagegen vorzugehen. Die Behörde behält sich zwei Monate vor, ihren eigenen Bescheid aufzuheben. Das Recht, der Automatik zu widersprechen, stand im Entwurf und wurde gestrichen. In Kraft ist das Gesetz noch nicht: Bundesrat und Kundmachung fehlen.

Parlament Österreich, Regierungsvorlage 539 d.B., Ausschussbericht 611 d.B., 96. Sitzung des Nationalrats, 23.09.2026 · Zum Vergleich: § 35a VwVfG (Deutschland), Art. 21 DSGVO (Schweiz).

## Dein Einkommen

Wovon lebst du, wenn eine Maschine deine Arbeit macht oder dein Kunde sie machen lässt?

## Deine Versorgung

Am 28. April 2025 fiel auf der Iberischen Halbinsel der Strom aus, 55 Millionen Menschen, rund zehn Stunden, ohne jeden Angriff. Wie viele Tage trägt dein Haushalt?

## Dein Staat

Was lässt du über dich ergehen, wenn eine Verwaltung unter Druck entscheidet, und niemand mehr zwischen dir und dem Bescheid steht?

## Dein Zugriff

Wer kann dir morgen Konto, Kommunikation und Dokumente sperren, und was tust du an diesem Tag?

Zwischen „Wir werden ausgelöscht“ und „Alle Krankheiten werden geheilt“ liegt dein Leben. Beide Ansagen kommen von denselben Leuten. Diese Lage erfasst jeder mit Hausverstand: ein Bauer, ein Installateur, eine Ärztin, ein Lehrer.

**Einordnung.** Resilienz ist nicht die Antwort auf die zwanzig Prozent. Gegen die Auslöschung baut niemand vor. Sie ist die Antwort auf die achtzig, über die keiner spricht.

# VIER AUSGÄNGE. AUF DREI IST NIEMAND VORBEREITET.

## Gegen uns

Ihre Ziele und unsere stehen gegeneinander. Befund 7 zeigt, wie das im Kleinen aussieht.

## Über uns

Die Systeme funktionieren, als Werkzeug derer, die sie kontrollieren. Überwachung schafft Untertanen.

## Für uns

Der versprochene Ausgang: Heilung, Wohlstand, Fortschritt. Möglich. Nur verlässt sich keiner der Erbauer allein darauf.

## An uns vorbei

Die Systeme verfolgen Ziele, in denen wir nicht vorkommen. Nicht Feindschaft. Gleichgültigkeit.

Welcher Ausgang eintritt, ist nicht belegbar. Belegbar ist, dass die Erbauer ihre Zahlen nennen und die Rechnung nicht zeigen. 1997 verlor der Mensch das Schachbrett und hat es nie zurückgewonnen. Die offene Frage ist, was er als Nächstes verliert.

## WAS DU JETZT TUN KANNST

- 1 Prüf es nach.** Das Kontrollverlust-Register führt jeden Vorfall mit Primärquelle und Datum: [ASlresilience.org/vorfaelle](https://ASlresilience.org/vorfaelle)
- 2 Stell die Frage, die keiner stellt.** Der Aufruf „Zeigt die Rechnung“ verlangt von den Erbauern die Aufteilung ihrer Zahlen. Mitzeichnen auf [ASlresilience.org](https://ASlresilience.org)
- 3 Sieh, wo es dich trifft.** Die Kontrollverlust-Analyse zeigt in zwei Minuten, welche Stellen deiner Lage offen sind: [ASlresilience.org/analyse](https://ASlresilience.org/analyse)
- 4 Geh in die Tiefe.** Das Buch: „Der KI-Kontrollverlust. Was dir über die Risiken der Superintelligenz bisher kaum jemand erklärt, welche Szenarien dich treffen können und wie du dich darauf vorbereitest.“ Richard Frederic Bertossa. [ASlresilience.org/buch](https://ASlresilience.org/buch)
- 5 Gib diesen Bericht weiter.** Unverändert und kostenlos. Das Thema läuft als Randnotiz, und es ist das existenzielle unter allen.