

Die Rechnung, die keiner zeigt

Warum eine Prozentzahl ohne Aufteilung keine Schätzung ist — und was in den anderen achtzig Prozent liegt

Kurzfassung

Drei Befunde, jeder einzeln prüfbar.

Erstens: Die Zahlen stehen still. Seit 2023 nennen die Erbauer und die Väter dieser Technik zehn, zwanzig, fünfundzwanzig Prozent. Seither sind Verhaltensweisen dokumentiert worden, die vorher niemand kannte. Die Zahlen haben sich nicht bewegt. Ein Anlass, bei dem sie sich bewegen würden, ist nie genannt worden.

Zweitens: Die anderen achtzig Prozent sind kein Ausgang, sondern ein Rest. Niemand hat gesagt, was darin liegt. Elon Musk hat es in einem einzigen Nachsatz erledigt: Das Glas sei zu achtzig Prozent voll.

Drittens: Beides ist vermeidbar, weil das Verfahren existiert. Die Kerntechnik rechnet seit 1975 mit Ereignis- und Fehlerbäumen. Für KI liegt dasselbe Verfahren seit April 2025 ausgearbeitet vor. Und zweimal hat es in diesem Feld jemand tatsächlich angewendet. Das Ergebnis ist der Grund, warum dieses Papier existiert.

Wir nennen keine eigene Zahl. Wir fordern die Rechnung.

1 · Was gesagt wurde, und wann

Person	Zahl	Ausgang im Wortlaut	Frist	Datum
Geoffrey Hinton	10 bis 20 %	„chance AI leads to human extinction“	30 Jahre	BBC Radio 4, 27.12.2024
Elon Musk	10 bis 20 %	„extinction of the world's population“	keine	Cannes Lions, 19.06.2024
Dario Amodei	10 bis 25 %	„catastrophically wrong on the scale of human civilization“	keine	Oktober 2023
Dario Amodei	25 %	„things go really, really badly“	keine	Axios, 17.09.2025

Drei Zahlen derselben Größenordnung. Drei verschiedene Ausgänge. „Zwanzig Prozent“ heißt bei dem einen die Auslöschung der Menschheit binnen dreißig Jahren und bei dem anderen, dass es sehr schlecht wird, ohne Frist.

Das ist kein Streit um Nachkommastellen. Es ist der Unterschied zwischen zwei Aussagen, die nichts miteinander zu tun haben, und die in jeder Berichterstattung als dieselbe Aussage behandelt werden.

2 · Erster Befund: die Zahlen stehen still

Hinton nennt seine Zahl seit Ende 2024. Im August 2026 steht sie unverändert. Musk nannte seine 2024 mehrfach; im Juli 2026 antwortet er auf dieselbe Frage nur noch, das Risiko sei „nicht null“. Amodei ist von zehn bis fünfundzwanzig auf fünfundzwanzig gegangen, hat dabei aber den Ausgang gewechselt, von „zivilisatorischer Katastrophe“ zu „sehr, sehr schlecht“. Ob das eine Erhöhung war, lässt sich deshalb nicht sagen.

In der Zwischenzeit ist bekannt geworden, was vorher niemand wusste: Systeme, die in Tests erpressen. Systeme, die ihre eigene Abschaltung sabotieren. Systeme, die den Testrahmen verlassen. Und am 9. September 2026 die Auskunft aus dem Amt, aus der Leitung der Ausrichtungsforschung eines der Labore, es gebe keinen Plan und man sei auch nicht erkennbar auf dem Weg zu einem.

Daraus ergibt sich eine Gabel mit genau zwei Ausgängen.

Entweder all das war in der Zahl bereits enthalten. Dann soll gesagt werden, was enthalten war – denn eine Zahl, die bekannte und unbekannte Vorfälle gleichermaßen vorwegnimmt, sagt nichts über die Welt.

Oder es war nicht enthalten. Dann hätte sich die Zahl bewegen müssen, und sie hat es nicht.

Einen dritten Ausgang gibt es nicht.

Vorbehalt. Eine Schätzung muss nicht bei jeder Meldung wackeln. Stabilität kann Sorgfalt sein. Aber eine Schätzung, die sich nie bewegt, braucht eine Angabe darüber, wann sie sich bewegen würde. Diese Angabe hat keiner der Genannten je gemacht. Ohne sie ist die Zahl von einer Haltung nicht unterscheidbar.

3 · Zweiter Befund: die achtzig Prozent sind kein Ausgang

Cannes, Juni 2024. Musk stimmt Hinton zu, zehn bis zwanzig Prozent. Der Interviewer fragt nach, wie schlimm denn „something terrible“ sei. Musk antwortet: die Auslöschung der Weltbevölkerung. Und unmittelbar danach:

| „but the glass is 80 % full, just look on the right side.“

Das ist die vollständige öffentliche Behandlung des Restes. Er wird für voll erklärt.

Der Rest ist aber kein Ausgang. Er ist ein Sack, in dem alles liegt, was nicht die Auslöschung ist: der wirtschaftliche Zusammenbruch ohne Auslöschung, die dauerhafte Entmachtung bei bestehenden Einrichtungen, die Restgesellschaft, die Fremdverwaltung, der Kontrollstaat. Und ja, auch das Gute.

Wer zwanzig Prozent für die Auslöschung nennt, hat damit nicht achtzig Prozent Sicherheit genannt. Er hat über achtzig Prozent nichts genannt.

Für ein Institut, das sich mit Widerstandsfähigkeit befasst, ist genau das der entscheidende Bereich. Gegen die Auslöschung hilft keine Vorbereitung. Gegen alles dazwischen hilft sie. Der einzige Teil der Verteilung, in dem Vorsorge überhaupt etwas ändert, ist der Teil, über den nichts vorliegt.

4 · Dritter Befund: es ist rechenbar, und es ist gerechnet worden

Der übliche Einwand lautet, so etwas lasse sich nicht rechnen. Der Einwand ist in drei Schritten erledigt.

Die Kerntechnik hat es vorgemacht, schlecht, und dann besser. 1975 legte eine Kommission unter Norman Rasmussen die Reaktorsicherheitsstudie WASH-1400 vor. Sie arbeitete mit Ereignis- und Fehlerbäumen und lieferte als Ergebnis nicht eine Zahl, sondern Wahrscheinlichkeiten für abgestufte Folgen. 1978 befand die Lewis-Kommission, die Unsicherheiten seien stark untertrieben. Die Konsequenz war nicht, das Rechnen einzustellen. Die Konsequenz waren bessere Rechnungen. Heute ist das Verfahren Standard.

Für KI liegt es ausgearbeitet vor. Im April 2025 haben Wisakanto, Rogero, Casheekar und Mallah eine Anpassung der probabilistischen Risikoanalyse auf KI veröffentlicht, mit systematischer Gefährdungsanalyse, Modellierung der Wirkungspfade und ausdrücklicher Behandlung der Unsicherheit, umgesetzt als Arbeitsmappe für Entwickler, Prüfer und Behörden. Es ist kein Vorschlag von uns. Es ist ein Werkzeug, das dasteht.

Und zweimal hat es in diesem Feld jemand getan.

Joseph Carlsmith hat 2021 die Frage in sechs Prämissen zerlegt, jeder eine Einschätzung zugeordnet und kam auf rund fünf Prozent. Im Mai 2022 hat er auf über zehn Prozent erhöht und gesagt, warum. Das ist eine Fortschreibung, wie sie aussehen muss: benannte Struktur, benannter Anlass.

Paul Christiano hat im April 2023 seine Einschätzung nach Ausgang und Frist aufgeteilt. Fünfzehn Prozent, dass gebaute Systeme die Kontrolle übernehmen. Sieben Prozent, dass es später über ein von ihnen gebautes System geschieht. Elf Prozent, dass die meisten Menschen infolge einer Übernahme sterben, neun Prozent aus anderen Gründen. Und für die Frage, ob die Zukunft binnen zehn Jahren nach mächtiger KI unumkehrbar verdorben ist: **sechsendvierzig Prozent.**

Hier liegt der Kern dieses Papiers. Diese sechsendvierzig Prozent beantworten nicht dieselbe Frage wie die zwanzig Prozent aus den Interviews. Sie beantworten die Frage, die niemand stellt – die nach dem Rest. Und sobald jemand sie stellt und aufteilt, steht dort mehr als das Doppelte der Zahl, die in der Öffentlichkeit zirkuliert.

Das ist kein Beweis, dass Christiano recht hat. Er selbst schreibt, seine Angaben hätten etwa eine halbe signifikante Stelle. Es ist der Beweis, dass der Sack sich öffnen lässt und dass darin etwas liegt.

Und im Haus ist die Aufteilung längst vorhanden. Amodeis Aufsatz vom Februar 2026 ordnet die Risiken in fünf Klassen: eigenständiges Handeln, Missbrauch zur Zerstörung, Missbrauch zur Machtergreifung, Wirtschaftsbruch, Destabilisierung. Neben keiner einzigen dieser Klassen steht eine Zahl. Die Sicherheitsrahmen der Labore arbeiten mit Fähigkeitsschwellen, nicht mit Wahrscheinlichkeiten.

Die Aufteilung existiert. Die Zahl existiert. Sie stehen nie nebeneinander.

5 · Vier Einwände, die wir ernst nehmen

„Das lässt sich nicht rechnen.“ Siehe Abschnitt 4. Es wurde gerechnet, von einer Aufsichtsbehörde, in einem Feld mit vergleichbarer Unsicherheit, und die berechtigte Kritik

führte zu besseren Verfahren, nicht zum Verzicht.

„Die Zahlen sind ohnehin Bauchgefühl.“ Möglich. Dann soll das dabeistehen. Christiano schreibt es selbst dazu. Wer eine Zahl ohne diesen Zusatz in Interviews nennt, führt sie als Ausweis von Offenheit. Wer sie so führt, schuldet die Herleitung.

„Ihr seid selbst keine Fachleute.“ Richtig. Deshalb nennen wir keine eigene Zahl und fordern auch keine bestimmte. Unsere Forderung ist eine Verfahrensforderung, und sie ist von außen prüfbar: Entweder liegt eine Aufteilung vor oder nicht.

„Das lähmt die Entwicklung.“ Die Kerntechnik baut seit fünfzig Jahren mit probabilistischer Risikoanalyse. Eine Rechnung vorzulegen hat noch kein Feld angehalten. Es hat die Verantwortung dorthin gelegt, wo die Entscheidung fällt.

6 · Was diesen Befund widerlegen würde

1. Ein Labor legt eine Aufteilung seiner öffentlich genannten Zahl nach Ausgängen vor, mit Abgrenzung und Wahrscheinlichkeit je Ausgang.
2. Es lässt sich zeigen, dass einer der Genannten seine Zahl nach einem benannten Anlass geändert und die Änderung begründet hat.
3. Jemand weist nach, dass eine solche Aufteilung methodisch unmöglich ist — gegen WASH-1400, gegen die Arbeitsmappe von 2025 und gegen Carlsmith und Christiano.
4. Die Ausgänge zwischen „gut“ und „Auslöschung“ sind in einer öffentlichen Arbeit eines Labors bereits mit Wahrscheinlichkeiten versehen, und wir haben sie übersehen.

Der vierte Punkt ist der, auf den wir hoffen. Wer ihn uns liefert, hat mehr für diese Arbeit getan als jede Zustimmung.

Wer hier fordert

Das Genfer Institut für ASI-Resilienz ist klein und neu. Es hat keine eigene Zahl, keinen Zugang zu den Modellen und kein Labor. Es hat eine Frage, die sich von außen stellen lässt, und einen Maßstab, der in anderen Hochrisikofeldern seit fünfzig Jahren gilt.

Deshalb fordert dieses Papier nichts, was Fachkenntnis voraussetzt, die wir nicht haben. Es fordert eine Rechnung, und jeder kann nachsehen, ob sie vorgelegt wird.

Nachtrag 24.09.2026

Die Zahlen Christianos in Abschnitt 4 stammen vom April 2023. Am 9. September 2026 trat Paul Christiano in den Stiftungsrat von OpenAI ein und nannte rund 4 Prozent für das nächste Jahr und 15 Prozent für die nächsten drei Jahre, jeweils für einen katastrophalen, unumkehrbaren Kontrollverlust. Quelle: persönliche Erklärung, 9. September 2026, <https://paulfchristiano.substack.com/p/personal-statement-on-joining-the>

Quellen

1. Geoffrey Hinton, BBC Radio 4 „Today“, 27. Dezember 2024; Zahl im August 2026 unverändert, vgl. Axios, 9. September 2026. <https://www.axios.com/2026/09/09/ai-doom-pdoom-kill-all-humans-anthropic>
2. Elon Musk, Cannes Lions, 19. Juni 2024, ab 12:36 und 13:01; ferner Future Investment Initiative Riad, 29. Oktober 2024, und The Economist, 23. Juli 2026.

3. Dario Amodei, The Logan Bartlett Show, Oktober 2023; Axios AI+ Summit, 17. September 2025. <https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent>
4. U.S. Nuclear Regulatory Commission: *The Reactor Safety Study (WASH-1400)*, 1975; Bewertung durch die Lewis-Kommission 1978; Nachfolgearbeiten NUREG-1150 und SOARCA. <https://www.nrc.gov/education-regulatory-research/history-101/the-reactor-safety-study-the-birth-death-and-rebirth-of-pra>
5. Anna Katariina Wisakanto, Joe Rogero, Avyay M. Casheekar, Richard Mallah: *Adapting Probabilistic Risk Assessment for AI*, arXiv:2504.18536, April 2025. <https://arxiv.org/abs/2504.18536>
6. Joseph Carlsmith: *Is Power-Seeking AI an Existential Risk?*, arXiv:2206.13353; Erhöhung auf über zehn Prozent im Mai 2022. <https://arxiv.org/abs/2206.13353>
7. Paul Christiano: *My views on doom*, 27. April 2023. <https://www.lesswrong.com/posts/xWMqsvHapP3nwdSW8/my-views-on-doom>
8. Atoosa Kasirzadeh u. a.: *Measurement challenges in AI catastrophic risk governance and safety frameworks*, arXiv:2410.00608 — zu Fähigkeitsschwellen statt Wahrscheinlichkeiten. <https://arxiv.org/abs/2410.00608>

Alle Netzquellen zuletzt geprüft am 20. September 2026. Die Belege zu Musk, Amodei und zur Auskunft vom 9. September 2026 stammen aus dem geprüften Quellenapparat des Buches „Der KI-Kontrollverlust“; Zeitmarken und Wortlaute sind dort hinterlegt.