

Christianos Rechnung

Die einzige öffentlich aufgeteilte Risikozahl aus dem Inneren des Feldes — und was am 9. September 2026 dazugekommen ist

WER DAS IST

Paul Christiano hat die Methode miterfunden, mit der heutige Sprachmodelle überhaupt brauchbar gemacht werden: das Lernen aus menschlicher Rückmeldung, RLHF, Grundlagenarbeit von 2017. Er leitete bei OpenAI das Team für Ausrichtung, verließ das Haus 2021 und gründete das Alignment Research Center. Seit April 2024 ist er Sicherheitschef der US-Behörde, die heute Center for AI Standards and Innovation heißt. Er verkauft kein Modell.

Am 27. April 2023 hat er seine eigene Einschätzung öffentlich aufgeteilt. Es ist keine Studie, kein Verfahren, keine Behörde. Es ist ein Mann, der seine Zahl zerlegt und jede Teilzahl benennt. Genau das hat kein Laborchef je getan.

WAS VOR ZWÖLF TAGEN GESCHEHEN IST

Am 9. September 2026 ist Christiano in den Stiftungsrat von OpenAI eingetreten, in den Ausschuss für Sicherheit, der das letzte Wort über die Freigabe neuer Modelle hat. Am selben Tag hat er eine persönliche Erklärung veröffentlicht. Darin stehen neue Zahlen, und diesmal mit Kalender.

April 2023	9. September 2026
Schnitt: Übernahme · Massensterben · unumkehrbar verdorben Bezug: zehn Jahre nach dem Bau mächtiger KI Ergebnis: 46 % unumkehrbar verdorben. Kein Datum.	Schnitt: katastrophaler, unumkehrbarer Kontrollverlust Bezug: das nächste Jahr, die nächsten drei Jahre Ergebnis: rund 4 % und 15 %. Mit Datum.

Derselbe Mann, drei Jahre Abstand, zwei verschiedene Schnitte. Das ist kein Vorwurf an ihn — es ist der Grund, warum ein gemeinsamer Maßstab fehlt.

Seine Worte: Das Risiko eines katastrophalen und unumkehrbaren Kontrollverlusts liege bei „roughly 4 percent over the next year and 15 percent over the next three years“. Und dazu: „If that happens then most people could die.“

Zwei weitere Sätze aus derselben Erklärung gehören dazu. Er rechnet damit, dass die KI-Forschung sich selbst automatisiert, „anywhere from several months to several years“. Und über die Lage der Branche, in deren Aufsichtsgremium er gerade eintritt: **„I do not think that the AI industry in general, including OpenAI, is currently on track to reduce this risk to an acceptable level.“**

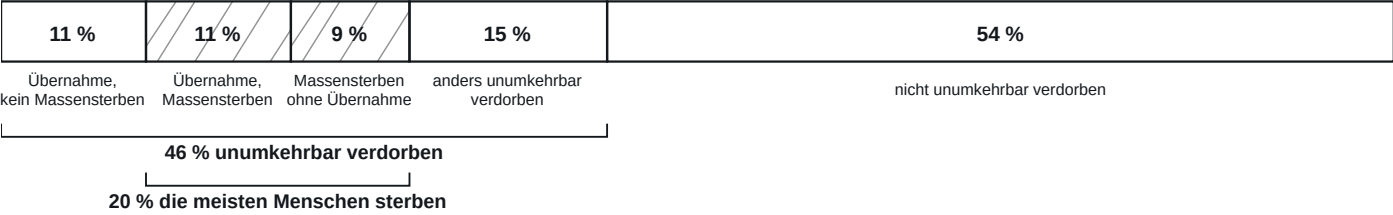
DIE RECHNUNG

WAS ÖFFENTLICH ZIRKULIERT

20 %	80 %
------	------

Auslöschung ein Rest ohne Aussage

CHRISTIANOS SCHNITT · 27. APRIL 2023



Die Übernahme macht zusammen 22 % aus: 15 % durch das gebaute System, 7 % später durch ein von ihm gebautes.

WAS AUCH BEI IHM NICHT AUFGETEILT IST

Wie sieht das Leben in diesen 46 % aus?	Und in diesen 54 %?
---	---------------------

Christiano teilt die erste Ebene auf: nach Mechanismus und nach Schwere. Die zweite Ebene — wie der Alltag in den Feldern aussieht — teilt auch er nicht auf.

ZAHL	WOFÜR SIE GILT, IN SEINEN WORTEN
22 %	Wahrscheinlichkeit einer Übernahme durch KI insgesamt
15 %	dass die von uns gebauten Systeme selbst übernehmen
7 %	dass diese nicht übernehmen, aber ein von ihnen gebautes, klügeres System es später tut
20 %	dass die meisten Menschen binnen zehn Jahren nach dem Bau mächtiger KI sterben
11 %	davon infolge einer Übernahme
9 %	davon aus anderen Gründen, als Folge des Baus oder im Umbruch danach
15 %	dass es auf andere Weise im beschleunigten Umbruch verdorben wird
46 %	dass die Menschheit ihre Zukunft binnen zehn Jahren nach dem Bau mächtiger KI unumkehrbar verdorben hat

WAS SIE BEANTWORTET — UND WAS NICHT

Sie beantwortet die Frage, ob hinter der Zahl etwas liegt. Ja. Wer den Sack öffnet, findet darin sechsundvierzig Prozent, wo außen zwanzig steht. Nicht weil er pessimistischer wäre, sondern weil er die Ausgänge zählt, die nicht das Ende sind.

Drei Dinge blieben offen. Eines davon hat er inzwischen selbst geschlossen, zwei nicht.

- 1. **Die zweite Ebene fehlt.** „Unumkehrbar verdorben“ ist selbst ein Sack, nur ein größerer. Fremdverwaltung, Kontrollstaat, Restgesellschaft, Entkopplung, Zusammenbruch ohne Auslöschung — alles liegt darin nebeneinander, ohne Trennung und ohne Zahl. Auch die vierundfünfzig Prozent sind nicht aufgeteilt.
- 2. **Der Kalender fehlte — und er hat ihn nachgeliefert.** Die Zahlen von 2023 gelten ab dem Tag, an dem mächtige KI gebaut ist, plus zehn Jahre; wann dieser Tag ist, steht nicht darin. Die Zahlen vom 9. September 2026 nennen ein Jahr und drei Jahre. Damit ist das die erste datierte Risikozahl mit

benanntem Ausgang, die es in diesem Feld gibt. Sie beantwortet aber eine andere Frage als die von 2023, also lassen sich die beiden nicht verrechnen.

3. **Der Einzelne kommt nicht vor.** Es ist eine Verteilung über die Welt, nicht über Orte, Berufe, Haushalte. Wo du in diesen Feldern stehst, steht nicht darin.

WIE WIR SIE VERWENDEN

Als Existenzbeweis, nicht als unsere Zahl. Der Satz lautet: Es geht. Einer aus dem Feld hat es getan, veröffentlicht, mit Namen darunter. Die, die bauen, haben es nicht getan. Damit ist der Einwand „so etwas lässt sich nicht aufteilen“ erledigt, ohne dass wir selbst eine Zahl nennen müssen.

Als Anschluss für die eigene Arbeit. Was wir beitragen, ist keine dritte Zahl, sondern die zweite Ebene: die Felder benennen, abgrenzen und mit Zeigern versehen, die man selbst ablesen kann. Dort steht bisher nichts — bei ihm so wenig wie bei den Laboren.

Als lebenden Beleg für die dritte Forderung. Der Aufruf verlangt, dass eine Zahl fortgeschrieben wird und dass gesagt wird, was sie bewegt. Genau das ist am 9. September geschehen, von einem Mann, der seither im Sicherheitsausschuss von OpenAI sitzt. Die Forderung ist damit keine Erfindung von uns mehr. Sie ist eine Praxis, die es gibt und der die Laborchefs nicht folgen.

Nicht als Gegenzahl zu den zwanzig Prozent. Die sechsundvierzig und die zwanzig beantworten verschiedene Fragen. Wer sie nebeneinanderstellt, ohne das zu sagen, macht denselben Fehler, den wir anprangern.

VORBEHALTE, DIE MITLAUFEN MÜSSEN

Er selbst schreibt dazu, es seien Schätzungen mit etwa einer halben signifikanten Stelle, die von Tag zu Tag schwanken — nicht wegen neuer Belege, sondern weil sie ungenau sind. Die Zahlen sind von 2023. Es ist eine Einzelmeinung, kein Verfahren, keine Behörde und keine Studie.

Und die Lücke, die wir benennen, hat vor uns schon jemand gesehen: In der Diskussion zu seinem Text im Mai 2023 wies ein Leser darauf hin, dass eine langlebige Diktatur unter KI weder ein guter Ausgang noch eine unumkehrbar verdorbene Zukunft wäre und in seinem Modell deshalb nirgends vorkommt. Neu an unserer Arbeit ist nicht die Beobachtung. Neu ist, daraus eine Forderung zu machen und sie an die Vorsorge zu binden.

Das ist kein Einwand gegen die Verwendung. Es ist der Grund für sie. Eine offengelegte Einzelschätzung mit Vorbehalt ist prüfbar. Eine Zahl ohne Herleitung ist es nicht.