

Im Internet geht gerade der folgenden Post viral (über 60 Millionen Klicks am 1.Tag)



Jacob Coxon ✓

@hilbertspaess

The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear privately. No other human activity poses this level of danger.

8:04 PM · Sep 8, 2026

Auf Deutsch;

Die Menschen, die KI entwickeln, glauben ernsthaft, dass sie uns alle bis zum Ende des Jahrzehnts töten könnte. Das ist kein Marketing-Gag. Wenn überhaupt, werden viele Führungskräfte und leitende Forscher ihre Formulierungen gegenüber der Presse abschwächen, um vernünftig zu wirken – aber privat höre ich dieselben Leute Angst äußern. Keine andere menschliche Tätigkeit birgt dieses Ausmaß an Gefahr.

8:04 Uhr · 8. September 2026

Jacob Coxon hat bei **Anthropic** gekündigt die kurz vor dem Börsengang steht. Wir sollten seine Warnung ernst nehmen!

Jacob Coxon der ehemaligen KI-Forscher von **OpenAI und Anthropic** –, hat sich gerade sehr kritisch zur weiteren Entwicklung von KI geäußert. Seine Hauptaussagen sind:

- **Die KI-Entwicklung geht seiner Meinung nach zu schnell.** Die großen KI-Labore würden in Richtung einer **selbstverbessernden Superintelligenz** laufen, also Systeme, die ihre eigenen Fähigkeiten immer weiter verbessern können.
- **Er hält diese Entwicklung für potenziell existenzbedrohend.** Coxon sagt, viele Menschen, die an KI arbeiten, glaubten privat tatsächlich, dass KI die Menschheit **noch vor Ende dieses Jahrzehnts** vernichten könnte.
- Er warnt davor, dass zukünftige KI-Systeme **übermenschliche Fähigkeiten** haben könnten – etwa beim Hacken, bei wissenschaftlicher Forschung oder beim Zugriff auf Ressourcen.

- Besonders gefährlich findet er den Moment, in dem KI **sich selbst verbessern kann**. Dann könnte die Entwicklung wesentlich schneller werden, als Menschen sie noch kontrollieren können.
- Er wirft **OpenAI und Anthropic** vor, in einem Wettlauf zu stehen: Beide könnten das Gefühl haben, dass sie selbst vorangehen müssen, weil sonst ein Konkurrent eine besonders mächtige KI entwickelt.
- Deshalb fordert Coxon **mehr Zusammenarbeit und eine Verlangsamung**. Er spricht unter anderem von Vereinbarungen zwischen den KI-Laboren, das Entwicklungstempo gemeinsam zu begrenzen – im Extremfall sogar von einem **zeitweisen Verbot weiterer Leistungssteigerungen**.

Was daran besonders interessant ist

Coxon sagt **nicht**, dass KI heute schon kurz davor ist, die Menschheit zu vernichten.

Seine Sorge ist vielmehr die **Entwicklungsrichtung**:

Je leistungsfähiger KI wird → desto autonomer wird sie → irgendwann könnte sie ihre eigene Weiterentwicklung beschleunigen → und dann könnte menschliche Kontrolle verloren gehen.

Das ist allerdings **Coxons Einschätzung, keine bewiesene Prognose**. Andere KI-Forscher halten solche Szenarien für weniger wahrscheinlich oder kritisieren, dass Coxon für einige seiner sehr weitreichenden Behauptungen keine konkreten Belege veröffentlicht hat.

Wie auch er würde sich dadurch strafbar machen.

Interessant ist aber: **Coxon ist nicht einfach ein Außenstehender**. Er hat drei Jahre an der Entwicklung bzw. am Vortraining großer KI-Modelle bei OpenAI und Anthropic gearbeitet. Deshalb wird seine Warnung innerhalb der KI-Branche durchaus ernst genommen.

