Eliezer Yudkowsky und Nate Soares: „If anyone builds it, everyone dies“

Wenn KI die Macht übernimmt

Von Vera Linß

Deutschlandfunk Kultur, Lesart, 19.09.2026

Etliche Fachleute warnen inzwischen vor der Auslöschung der Menschheit durch eine Künstliche Intelligenz. Die KI-Forscher Eliezer Yudkowsky und Nate Soares legen nun dar, wie sich das genau vollziehen könnte. Ihr oft als Alarmismus abgetanes Szenario überzeugt.

Die Nachrichten mehren sich. Ob bei OpenAI, Meta oder Anthropic: Immer wieder gelingt es Künstlichen Intelligenzen anscheinend, ein Eigenleben zu entwickeln und Aktivitäten zu entwickeln, für die sie nicht trainiert worden sind. Noch bekommen die KI-Unternehmen solche Entwicklungen in den Griff. Das allerdings dürfte bald vorbei sein, prophezeien die KI-Forscher Eliezer Yudkowsky und Nate Soares.
Sie fordern, die Entwicklung leistungsfähigerer KI-Algorithmen einzustellen. Anderenfalls drohe die Auslöschung der Menschheit durch Künstliche Intelligenz.

Eliezer Yudkowsky und Nate Soares

If anyone builds it, everyone dies

Übersetzt von Nikolas Bertheau

Droemer Verlag. München 2026

304 Seiten

18 Euro

Die beiden Wissenschaftler vom gemeinnützigen Machine Intelligence Research Institut in Kalifornien sind nicht die einzigen, die vor einer KI-Eskalation warnen. Schon 2023 wiesen etliche KI-Pioniere auf solche Gefahren hin. Und auch aktuell kocht die Debatte darüber wieder hoch.

Das fiktive Modell „Sable“

Doch wie wahrscheinlich ist diese Dystopie, die für viele nach Alarmismus klingt? Glaubt man Eliezer Yudkowsky und Nate Soares, tritt die feindliche Übernahme durch eine „Superintelligenz“ mit Sicherheit ein, wenn sie entwickelt wird. Das Besondere an ihrem Buch: Sie können ihre Befürchtung auch plausibel erklären.

Dabei behaupten sie nicht zu wissen, zu welchem Zeitpunkt eine selbständig werdende KI beginnen werde, sich gegen die Menschen zu richten. Und auch wie das genau ablaufen solle, können sie nicht sagen. Aber sie stellen ein Szenario vor, das einen möglichen Auslöschungsprozess anschaulich beschreibt. Dafür haben sie das fiktive Modell „Sable“ erfunden, mit dem sie stellvertretend durchspielen, mit welchen Mitteln es der KI gelingen könnte, heimlich eine Erweiterung ihrer selbst zu schaffen und mit dieser am Menschen vorbei zu operieren.

Beklemmend, wieviel Realitätsbezug in „Sable“ steckt. Wie es zunächst im Verborgenen Eigenarten entwickelt und diese Fähigkeiten nach und nach verstärkt. Um sich dann – sobald die Entwickler das Modell veröffentlicht haben – eigenständig in Unternehmensnetzwerken auszubreiten. Trickreich auch, wie „Sable“ an Geld für Computerchips kommt, um seine Hardware aufzustocken. Der Diebstahl von Kryptogeld oder die Erpressung von Menschen mit gestohlenen Daten – für „Sable“ kein Problem. Ob die KI dabei nach und nach Bewusstsein erlangt, ist für die Forscher allerdings nicht relevant. Ihnen reicht die Annahme, dass sie Intentionen entwickelt. Am Ende schafft es die KI, Forschungsarbeit zu tödlichen Viren auszulösen, gegen die die Menschheit machtlos ist.

Anders als andere Wissenschaftler, die solch ein Risiko – wenn überhaupt – eher als Kollateralschaden vermuten, gehen Eliezer Yudkowsky und Nate Soares davon aus, dass sich KI (auch) dezidiert gegen den Menschen richten würde. Denn neben konkurrierenden Algorithmen würden auch die Menschen versuchen, die KI einzuschränken: Störfaktoren, die es beseitigen gilt.

KI-Forschung auf dem Niveau der Alchimie

Ihr Szenario unterlegen Eliezer Yudkowsky und Nate Soares immer wieder mit gut verständlichen Erläuterungen darüber, wie das Training von Algorithmen technisch funktioniert und welche Wege es nehmen kann. Dabei erklären sie auch, warum sie einen Kontrollverlust erwarten. Ihre Prämisse: KI, die man zu Höchstleistungen trainiert, wird immer versuchen, ihre Ressourcen auszuschöpfen – auch darüber hinaus, was ein Mensch für sie vorgesehen hat. Um dem Nachdruck zu verleihen, sparen sie nicht an wortstarker Kritik an all den „Quacksalbern“, die das Gegenteil behaupten. Wenn es um die Funktionsprinzipien von (Künstlicher) Intelligenz gehe, bewege sich die Forschung auf dem Niveau der Alchimie.

Dass nur ein Entwicklungsstopp von KI – wie aktuell wieder besonders vehement gefordert – die „Katastrophe“ abwenden kann, ist da nur naheliegend. Auch wenn vieles dafür spricht, dass es den großen KI-Unternehmen wie OpenAI oder Anthropic dabei vor allem um ihre eigenen Geschäftsinteressen geht, bleibt das ungelöste Problem des „AI-Alignment“, also der Frage, wie man KI im Sinne des Menschen ausrichten kann, dennoch bestehen. Gelingen soll dies laut Yudkowsky und Soares mit einer konzertierten Aktion „der wichtigsten Mächte“ – Ländern wie die USA, Großbritannien, China oder Russland, vergleichbar mit Abkommen, wie sie es zur Kontrolle von Nuklearwaffen gibt. Auch wenn das kaum machbar erscheint: Dass man dies zügig angehen sollte, ist die wichtigste Botschaft dieses überzeugenden Buches.