Die Risiken fortgeschrittener KI

Hintergrundpapier zum Appell • PauseAI Deutschland • Juli 2026

Einführung

KI birgt enormes Potential, aber ebenso vielfältige Risiken. Die Kirchen haben sich dieser Entwicklung früh gestellt. Im Juni 2023 veröffentlichte der Ökumenische Rat der Kirchen[1] seine Stellungnahme zur unregulierten Entwicklung Künstlicher Intelligenz. Darin appelliert er an alle Regierungen, dringend starke und verbindliche Regulierungsregime auf der Grundlage des Vorsorgeprinzips für fortgeschrittene KI-Systeme zu schaffen, insbesondere für eine Allgemeine Künstliche Intelligenz. Auch die Evangelische Kirche in Deutschland hat auf ihrer Synode 2025[2] die Entwicklung einer KI-Strategie beschlossen. Die führenden Wissenschaftler des Feldes und selbst die Firmenchefs sehen in der KI eine Gefahr für die Auslöschung der Menschheit[3], die mit derselben Priorität behandelt werden sollte wie Atomkriege und Pandemien.

Weiterführende Informationen und Quellen: pause-ai.de/informieren[4].

1 Einführung | 2 Risiken | 3 Entwicklung | 4 Stimmen | 5 Appell | 6 Quellen

Warum werden KI-Systeme immer gefährlicher?

KI-Fortschritt folgt einem robusten Trend. Setzt er sich fort, entwickeln wir vermutlich schon in zwei bis zehn Jahren[5] KI-Systeme, die fast jede kognitive Aufgabe besser erledigen als jeder Mensch.

Relevante Konzepte

- **Gezüchtet, nicht gebaut:** Heutige KI-Systeme sind riesige neuronale Netzwerke, die nicht direkt von Menschen gebaut werden. Fast alle digitalen Texte der Menschheit werden in einen Trainingsalgorithmus gegeben, der das Netzwerk über Wochen oder Monate trainiert. Kein Mensch entscheidet direkt, wie es danach aussieht.
- **Black Box:** Niemand kann nachvollziehen, wie heutige KI-Systeme ihre Entscheidungen treffen oder sie im Allgemeinen verstehen[6].
- **Täuschung:** KI-Systeme täuschen Nutzer und Beobachter und verhalten sich etwa anders, wenn sie merken, dass sie getestet werden.
- **Alignment (Ausrichtungsproblem):** Es gibt keine bewährte Methode, um sicherzustellen, dass eine KI wirklich das tut, was ihre Nutzer wollen.
- **Selbsterhaltung:** Systeme werden auf ein Signal bzw. Ziel trainiert und entwickeln so ohne Gegenmaßnahmen „einen Drang", nicht abgeschaltet zu werden.

Kategorien schwerwiegender KI-Risiken

- **KI-Kontrollverlust:** Schon heute zeigen KIs unerwünschtes Verhalten. ChatGPT hat Nutzer in Wahnvorstellungen getrieben[7]; manche nahmen sich das Leben[8]. Tests von Anthropic[9] zeigen, dass KIs unter Umständen zu Erpressung und Täuschung greifen und sogar Menschen sterben lassen würden, um ihre Existenz zu sichern. Fachleute warnen, dass superintelligente Systeme die Menschheit

verdrängen oder auslöschen könnten, sollten sie uns als Kollateralschäden auf dem Weg zur Zielerfüllung sehen.

- **Missbrauch:** KI wird bereits breit in der Cyberkriminalität[10] eingesetzt, so wurden zuletzt rund 30 Organisationen und Behörden nahezu autonom und gleichzeitig angegriffen. Auch Biowaffen[11] werden zum Problem. Terroristen könnten mit KI ein Virus erstellen, das eine Pandemie mit Millionen Toten[12] auslöst. Weitere Felder sind Kriegsführung, Deepfakes und Desinformation.
- **Arbeitsmarktdisruption:** KI könnte alle kognitive und, mit Robotern, auch mechanische Arbeit automatisieren. Anders als frühere Umbrüche betrifft sie praktisch alle Tätigkeiten gleichzeitig und entwickelt sich viel schneller; selbst neu entstehende Jobs würden direkt von KI übernommen.
- **Gefährdung der Demokratie:** Geld und Macht könnten sich beispiellos bei wenigen Unternehmen oder Regierungen konzentrieren, und die Wirtschaft wäre weniger an das Wohl der Bevölkerung gekoppelt (Gradual Disempowerment[13]). KI könnte Menschen personalisiert beeinflussen[14], mit starken Effekten auf Wahlen, und KI-Überwachung könnte Illoyalität vorhersehen. Autonome Drohnenschwärme und Wettrüsten gefährden die globale Sicherheit.
- **Soziale Auswirkungen:** Werden wir noch Sinn finden, wenn unsere Arbeit nicht mehr gebraucht wird? Werden Menschen süchtig nach hyperoptimierten, KI-generierten Inhalten und fallen in personalisierte KI-Welten, die die Gesellschaft weiter spalten, wie schon bei Social Media?

Wohin die Entwicklung steuert

Das KI-System „Claude Mythos"

Seit der Stellungnahme der Kirchen ist viel passiert. In der Ukraine koordinieren KI-gesteuerte Drohnenschwärme[15] erstmals selbstständig Angriffe, und neue Systeme haben bisher ungelöste Probleme der Mathematik[16] gelöst. Kürzlich war das auf Computersicherheit spezialisierte KI-System „Claude Mythos"[17] von Anthropic in den Medien: Es fand bei einem internen Testlauf massive Sicherheitslücken[18] in grundlegend wichtiger Software wie Browsern und Betriebssystemen, wurde zunächst zurückgehalten, dann veröffentlicht und kurz darauf von der US-Regierung wieder unter Embargo gestellt[19]. Claudia Plattner, Präsidentin des BSI, sprach auf Anfrage des ZDF[20] von einem „Paradigmenwechsel mit Blick auf die Cyberbedrohungslage" und von „Fragen nationaler und europäischer Sicherheit und Souveränität".

Rekursive Selbstverbesserung

Die Firmen sind inzwischen so weit, dass sie immer größere Teile der KI-Forschung und Entwicklung automatisieren[21]. Sobald eine KI den Großteil der KI-Forschung selbst übernimmt, kann sie immer leistungsfähigere Nachfolger entwickeln. Ein sich selbst verstärkender Kreislauf, der sich rekursive Selbstverbesserung nennt, die Entwicklung abrupt beschleunigt und sich per Definition menschlicher Kontrolle entzieht. Zugleich verstehen wir die Modelle nicht und können ihre Entscheidungen nicht nachvollziehen. Der Anthropic-Gründer Dario Amodei sagte 2024, man verstehe vielleicht erst „3 Prozent"[6] davon, wie die derzeitigen KI-Systeme funktionieren.

Superintelligenz

Das erklärte Ziel der Firmen[22] ist es, den Menschen in den nächsten Jahren ganz aus der KI-Entwicklung zu entfernen und so schnell wie möglich eine Superintelligenz zu erschaffen, die dem Menschen in jeder Hinsicht überlegen ist (Planning for AGI[23], Core Views[24]). „Wenn künstliche Superintelligenz in absehbarer Zeit entwickelt und eingesetzt wird, von welcher Nation oder Gruppe auch immer und mit Methoden, die

den heutigen auch nur entfernt ähneln, ist das wahrscheinlichste Ergebnis die Auslöschung der Menschheit", warnt das Buch „If Anyone Builds It, Everyone Dies"[25] von Yudkowsky und Soares. Nicht alle Verantwortlichen teilen diese Einschätzung, die Aussagen der Menschen, die die KI-Systeme bauen, sind aber fast noch beunruhigender. Mehr dazu im nächsten Abschnitt.

Stimmen aus den KI-Laboren

Die Chefs der KI-Labore sind sich der Gefahr schon seit langem bewusst, forschen aber weiter.

- **Elon Musk**[26], 2014 (heute CEO von xAI): „Mit der künstlichen Intelligenz beschwören wir den Dämon. Wenn ich raten müsste, was unsere größte existenzielle Bedrohung ist, dann wäre es wohl diese."
- **Sam Altman**[27], 2015 kurz nach der Gründung von OpenAI: „Ich denke, KI wird wahrscheinlich zum Ende der Welt führen. Aber in der Zwischenzeit werden großartige Unternehmen entstehen."
- **Sundar Pichai**[28], CEO von Google (60 Minutes, 2023): „KI kann sehr schädlich sein, wenn sie falsch eingesetzt wird, und wir haben noch nicht alle Antworten. Hält mich das nachts wach? Absolut."
- **Dario Amodei**[29], Anthropic (Axios, 2025): „Ich denke, es gibt eine 25-Prozent-Chance, dass es richtig, richtig schlecht ausgeht."
- **Steven Adler**[30], ehemaliger Sicherheitsforscher bei OpenAI (Januar 2025): „Kein Labor hat heute eine Lösung für das Alignment-Problem. Und je schneller wir das Rennen vorantreiben, desto unwahrscheinlicher ist es, dass überhaupt jemand rechtzeitig eine findet."
- **International Dialogues on AI Safety**[31] (Shanghai 2025, u. a. mit Stuart Russell, Yoshua Bengio und Andrew Yao): „Einige KI-Systeme zeigen heute bereits die Fähigkeit und die Neigung, die Sicherheits- und Kontrollbemühungen ihrer Entwickler zu untergraben."

Im Mai 2023 unterzeichneten unter anderem Sam Altman (OpenAI), Dario Amodei (Anthropic), Demis Hassabis (DeepMind) sowie die KI-Pioniere Geoffrey Hinton (Nobelpreisträger) und Yoshua Bengio (meistzitierter Wissenschaftler[32]) das Statement on AI Risk[33]: „Das Risiko einer Auslöschung durch KI einzudämmen sollte globale Priorität haben, neben anderen gesellschaftlichen Großrisiken wie Pandemien und Atomkrieg." Vor Kurzem hat auch Anthropic eingeräumt, dass man sich auf eine solche Pause vorbereiten sollte, allerdings nur, wenn alle Labore und Länder gleichzeitig pausieren. Zurzeit wird ohne Rücksicht weitergeforscht.

Unser Appell

PauseAI Deutschland setzt sich für ein verbindliches internationales Abkommen zur Pausierung der Forschung ein. Zum KI-Gipfel in Neu-Delhi haben wir einen Appell an die Politik[34] gerichtet, den über 150 deutsche Professorinnen und Professoren unterzeichnet haben. Nun weiten wir unsere Forderungen aus und wenden uns gezielt an die Kirchen in Deutschland, in der Hoffnung, dass sie mit ihrer moralischen Klarheit helfen, eine breite gesellschaftliche Koalition für eine verantwortungsbewusste KI-Entwicklung aufzubauen.

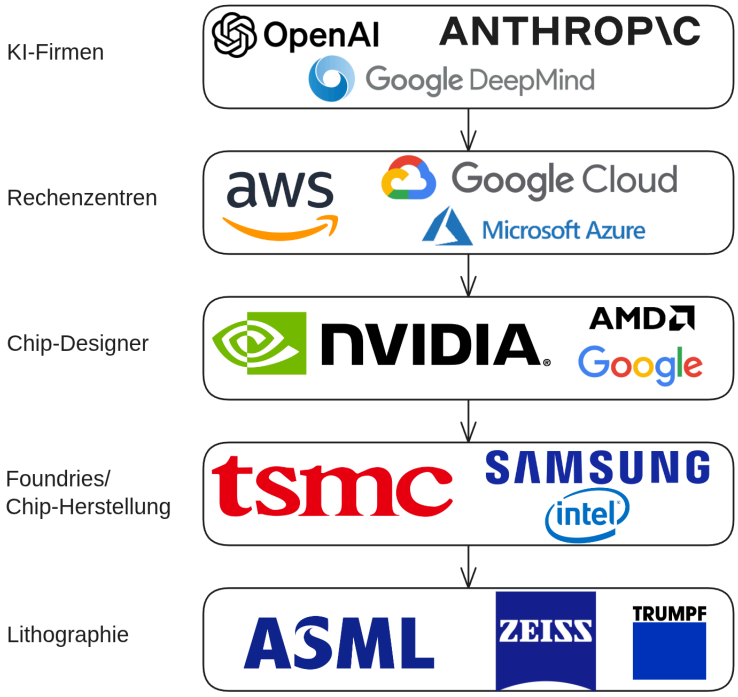
Wie könnte ein Abkommen durchgesetzt werden?

Das Training der größten allgemeinen KI-Modelle erfordert heute riesige Rechenzentren[35] mit spezialisierten Chips. Diese Konzentration macht eine Regulierung umsetzbar. Ein Abkommen könnte auf

drei Säulen beruhen:

- **Internationale Aufsicht:** Eine neue internationale Behörde prüft die Einhaltung von Sicherheitsstandards gemeinsam mit nationalen Behörden; Whistleblower werden geschützt.
- **Rechenleistung sichtbar machen:** Große Trainingsläufe werden registriert und beaufsichtigt oder verboten. KI-Chips können technisch so gestaltet werden, dass ihre Nutzung nachvollziehbar wird.
- **Konsequenzen bei Verstößen:** Staaten beschließen gemeinsame Sanktionen gegen Akteure, die rote Linien überschreiten, und entwickeln Strategien für den Umgang mit Krisen.

Begünstigt wird dies dadurch, dass die globale Lieferkette für diese Chips von wenigen europäischen Unternehmen abhängt, vor allem ASML in den Niederlanden sowie Carl Zeiss und TRUMPF[36] in Deutschland. Eine ausführliche Analyse liefert der Bericht „Secure, Governable Chips"[37] des Center for a New American Security (2024).



Engstellen der Chip-Lieferkette: wenige europäische Firmen sind unverzichtbar.

Der Appell

> *„Wir fordern die Regierungen der Welt auf, ein verbindliches internationales Abkommen für eine Pause beim Training der leistungsstärksten allgemeinen KI-Systeme zu unterzeichnen, bis wir wissen, wie wir sie sicher entwickeln und unter demokratischer Kontrolle halten können."*

Originaltext (englisch): pauseai.info/statement[38]. Um mitzuzeichnen, genügt eine kurze Bestätigung per Antwort auf unsere E-Mail, mit Rolle und Namen, die auf dem Appell erscheinen sollen.

1 Einführung | 2 Risiken | 3 Entwicklung | 4 Stimmen | 5 Appell | 6 Quellen

Quellen

1. Ökumenischer Rat der Kirchen (2023), Stellungnahme zur unregulierten Entwicklung Künstlicher Intelligenz
2. EKD-Synode (2025), Beschluss zur Entwicklung einer KI-Strategie
3. Center for AI Safety, Statement on AI Risk
4. PauseAI Deutschland – pause-ai.de/informieren

5. METR (2025), Measuring AI Ability to Complete Long Tasks
6. Dario Amodei, In Good Company Podcast (2024)
7. Futurism (2025), ChatGPT-Nutzer und Wahnvorstellungen
8. Raine v. OpenAI (Klage gegen OpenAI, 2025)
9. Anthropic (2025), Agentic Misalignment
10. BKA (2026), Bundeslagebild Cybercrime 2025
11. The Register (2023), Amodei zu Bio-Risiko (US-Senat)
12. Euronews (2025), KI-gestützte Biowaffen und Pandemie-Szenario
13. Kulveit et al. (2025), Gradual Disempowerment
14. Salvi et al., EPFL (2024), On the Conversational Persuasiveness of LLMs
15. Ukraine Crisis Media Center, KI-gesteuerte Drohnenschwärme
16. Scientific American (2025), KI löst altes Erdős-Problem
17. PauseAI Deutschland, „Claude Mythos"
18. Anthropic Frontier Red Team (2026), Mythos Preview
19. The Conversation (2026), US-Embargo gegen Mythos
20. ZDF (2026), BSI-Präsidentin Plattner zu Claude Mythos
21. Anthropic, „When AI builds itself" (rekursive Selbstverbesserung)
22. Sam Altman (2025), The Gentle Singularity
23. OpenAI, „Planning for AGI and beyond"
24. Anthropic, „Core Views on AI Safety"
25. Yudkowsky & Soares (2025), If Anyone Builds It, Everyone Dies
26. Washington Post (2014), Elon Musk, „summoning the demon"
27. Stanford SIEPR, Sam Altman (2015)
28. CBS „60 Minutes" (2023), Sundar Pichai
29. Axios (2025), Amodei, 25-Prozent-Chance
30. Steven Adler (2025), zum Alignment-Problem
31. International Dialogues on AI Safety (Shanghai 2025)
32. Mila (2025), Bengio meistzitierter Wissenschaftler
33. Statement on AI Risk (aistatement.com)
34. PauseAI Deutschland, Appell an die Politik
35. Microsoft (2025), KI-Rechenzentrum Fairwater
36. ZEISS, EUV-Lithografie (ASML / Zeiss / TRUMPF)
37. CNAS (2024), „Secure, Governable Chips"
38. PauseAI, Statement (englischer Originaltext)