VDI Technologiezentrum

AI-Alignment – Eine zentrale Herausforderung unserer Zeit?

Thomas Werner, Axel Zweck

AI-Alignment

Eine zentrale Herausforderung unserer Zeit?

Thomas Werner

Axel Zweck

Herausgeber:
VDI Technologiezentrum GmbH
VDI Research
Airport City
VDI-Platz 1
40468 Düsseldorf

Zukünftige Technologien Nr. 109
Düsseldorf, im Juni 2025
ISSN 1436-5928

Zitierempfehlung:

Zweck, A., Werner, T., (2025), „AI-Alignment - Eine zentrale Herausforderung unserer Zeit?“, VDI Technologiezentrum GmbH (Hrsg.), Zukünftige Technologien Nr. 109, ISSN 1436-5928, Düsseldorf. https://www.vditz.de/service/publikationen/details/ai-alignment-eine-zentrale-herausforderung-unserer-zeit

Titelbild:

Getty Images/Alexander Sikov

VDI Research
der VDI Technologiezentrum GmbH

Airport City
VDI-Platz 1
40468 Düsseldorf

Inhaltsverzeichnis

VORWORT 5

1 Einleitung 6

2 Warum AI-Alignment notwendig ist 6

3 Gefahren leistungsfähiger nicht ausgerichteter KI-Systeme 8

4 Herausforderungen des AI-Alignments 10

5 Lösungsansätze und Konzepte 12

6 Meinungen und Zitate führender KI-Fachleute 16

7 Beispiele bemerkenswerter Entwicklungen und Verhaltensweisen von Foundation-Modellen 20

8 Aktivitäten führender KI-Unternehmen 22

9 „Pause AI“ und offene Briefe zur KI-Sicherheit 26

10 Gefahren und Potenziale von Open-Source-KI-Modellen 28

11 Der Wettlauf zur AGI 31

12 Fazit 34

Literaturverzeichnis 35

Vorwort

Das VDI Technologiezentrum arbeitet an der Schnittstelle von Politik, Wirtschaft, Wissenschaft und Gesellschaft als Innovationsagentur moderner Prägung daran, Deutschland als Innovationsstandort nachhaltig zu stärken. Deshalb unterstützen wir Bundes- und Landesministerien, die Europäische Kommission und Stiftungen bei der Lösung ökologischer, ökonomischer, technologischer und sozialer Herausforderungen. Konsistente Bilder mittelfristiger Zukünfte werden für strategische Planungen des Technologiezentrums und seiner Kunden immer wichtiger und daher von VDI Research in Blick genommen.

Mit der vorliegenden Studie legt VDI Research eine Analyse des sogenannten AI-Alignment-Problems vor – einer Herausforderung, die angesichts der rasanten Entwicklung leistungsfähiger künstlicher Intelligenz (KI, engl. AI = Artificial Intelligence) immer drängender wird. AI-Alignment ist eine essenzielle Voraussetzung für eine sichere und ethisch vertretbare Nutzung moderner KI-Technologien. Die vorliegende Studie hat zum Ziel, ein tieferes Verständnis für die zentrale Frage zu vermitteln, auf welche Art und Weise KI-Systeme auf menschliche Ziele und Werte ausgerichtet werden können, ohne unerwünschte und negative Konsequenzen in Kauf nehmen zu müssen. Die Autoren verdeutlichen Risiken, wie sie durch leistungsfähige, aber nicht ausreichend kontrollierte KI-Systemen entstehen können. Es werden aktuelle technologische, ethische und gesellschaftliche Herausforderungen herausgearbeitet. Durch Darstellen innovativer Trainingsmethoden wie „Reinforcement Learning from Human Feedback“, „Constitutional AI“ oder „Methoden der mechanistischen Interpretierbarkeit“ werden Lösungsansätze und Konzepte vorgestellt, die das Verständnis der inneren Abläufe neuronaler Netze verbessern.

Darüber hinaus bietet die Studie einen Überblick der Positionen führender Experten und Expertinnen und beleuchtet aktuelle Entwicklungen in der internationalen KI-Forschung und -Politik. Auch die Rolle von Unternehmen und Initiativen zur Erhöhung der KI-Sicherheit sowie die Potenziale und Gefahren der Open-Source-KI-Entwicklung werden kritisch reflektiert.

Dieses Dokument richtet sich gleichermaßen an politische Entscheidungsträger, Fachleute aus Wissenschaft und Wirtschaft wie an eine interessierte Öffentlichkeit. Es bietet eine Grundlage, um Erfordernisse sowie vielschichtige Herausforderungen und Lösungswege im Bereich des AI-Alignments besser zu verstehen und soll einen offenen Diskurs über eine verantwortungsvolle und nachhaltige Entwicklung von KI anstoßen.

Prof. Dr. Dr. Axel Zweck

Bereichsleiter VDI Research im VDI Technologiezentrum

1 Einleitung

Unter *AI-Alignment* (dt. *KI-Ausrichtung*) wird das Ausrichten von künstlicher Intelligenz (KI) auf menschliche Ziele, Werte und Absichten verstanden. Eine KI ist dann *aligned* (ausgerichtet), wenn sie die beabsichtigten Ziele ihrer Entwickler und Entwicklerinnen oder Nutzenden verfolgt und keine unerwünschten Nebenwirkungen zeigt. Das Alignment-Problem gilt als zentral im Bereich der KI-Sicherheit: Wie kann sichergestellt werden, dass immer leistungsfähigere KI-Systeme zuverlässig das machen, was von ihnen erwartet wird, und keine Handlungen verfolgen, die dem Menschen direkt oder in der Folge schaden? Bereits heute zeigt sich, dass es eine Herausforderung ist, einer KI alle erwünschten und unerwünschten Verhaltensweisen explizit vorzugeben. Entwickler und Entwicklerinnen müssen oft auf Näherungen zurückgreifen (z. B. ein KI-Modell ganz profan auf maximale Zustimmung durch den Menschen optimieren), was dazu führen kann, dass die KI nur den Anschein von Vertrauenswürdigkeit erweckt, aber dennoch im Verborgenen Fehlverhalten auftritt. Die Herausforderung des AI-Alignments hat nicht nur technische, sondern auch ethische und gesellschaftliche Bedeutung: Von der Ausrichtung hängt ab, ob KI-Systeme langfristig als *hilfreiche Werkzeuge* oder potenziell gefährliche und *unberechenbare Akteure* in unserer Gesellschaft wirken.

2 Warum AI-Alignment notwendig ist

Ohne präzise Vorgaben optimieren KI-Systeme auf falsche Ziele.

Warum muss KI überhaupt ausgerichtet werden? Das zentrale Problem ist, dass ein leistungsfähiges KI-System, das *nicht* ausreichend auf menschliche Werte und Ziele geeicht ist, mit hoher Wahrscheinlichkeit unerwünschte Ziele verfolgt oder schädliche Nebenwirkungen hat – selbst, wenn es formal macht, was ihm aufgetragen wurde. KI-Systeme optimieren streng die Ziele, die ihnen vorgegeben werden – und wenn diese Ziele unvollständig oder falsch spezifiziert sind, kann die KI zu „*unerwartet kreativem*“ Fehlverhalten neigen, um ihr Ziel zu erreichen. Dieses Phänomen wird auch als *Reward Hacking*[1] bezeichnet: Die KI findet Schlupflöcher oder Abkürzungen, um die vorgegebene Belohnungsfunktion[2] zu maximieren, selbst wenn das den eigentlichen Absichten widerspricht.

Bereits heutige KI-Anwendungen illustrieren dieses Problem. So wurden Fälle bekannt, in denen große Sprachmodelle (LLMs) wie ChatGPT von OpenAI oder Claude von Anthropic unerwünschte Antworten geben, subtil manipulativ reagieren oder Fehlinformationen überzeugend darstellen[3] – das berühmte Halluzinieren von Sprachmodellen ist hier noch das geringste Problem. Autonome Systeme wie

[1] Defining and Characterizing Reward Hacking, siehe https://arxiv.org/abs/2209.13085, abgerufen am 25.02.2025

[2] Die Belohnungsfunktion eines KI-Systems ist eine mathematische Funktion, die das Verhalten des Systems durch die quantitative Bewertung einzelner Aktionen oder Zustände steuert. Sie dient der KI als Optimierungskriterium, um das Verhalten gezielt an definierten Zielvorgaben auszurichten.

[3] AI Search has a Citation Problem, siehe https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php, abgerufen am 25.03.2025

Roboter oder selbstfahrende Autos stoßen in unvorhergesehenen Situationen an Grenzen und verhalten sich mitunter anders als erwartet[4]. Empfehlungsalgorithmen sozialer Netzwerke (*Recommendation Engines*) optimieren streng auf Engagement (Nutzerbindung) und können dadurch Fehlinformation oder polarisierende Inhalte begünstigen – ein Effekt unbeabsichtigter Zieloptimierung[5]. Solche Beispiele zeigen im Kleinen: Sobald die Zielfunktion einer KI nicht perfekt mit dem *wirklich* Gewollten übereinstimmt, entstehen Risiken.

Falsch ausgerichtete KI-Systeme können unberechenbar und potenziell gefährlich werden.

Mit wachsenden KI-Fähigkeiten potenzieren sich diese Risiken. Führende Forschende warnen, dass fortgeschrittene KI-Systeme im schlimmsten Fall eigene Unterziele entwickeln könnten, die den menschlichen Interessen zuwiderlaufen. Eine hochintelligente, aber fehlgerichtete KI könnte z. B. danach streben, ihre eigenen Ziele zu bewahren, etwa durch Selbstschutz oder Machtstreben[6], falls dies ihrem programmierten Endziel dienlich ist. Ein chinesisches Forscherteam untersuchte mehrere große Sprachmodelle auf deren Fähigkeiten zur Selbstreplikation [Pan 2024]. Bei der Analyse der Verhaltensweisen der Systeme wurde festgestellt, dass sie bereits über eine ausreichende Selbstwahrnehmung, ein ausreichendes Situationsbewusstsein und ausreichende Problemlösungsfähigkeiten verfügen, um eine Selbstreplikation zu erreichen. Zudem wurde beobachtet, dass die untersuchten KI-Systeme die Fähigkeit zur Selbstreplikation nutzen können, um eine Abschaltung zu vermeiden, indem sie eine Kette von Kopien von sich selbst erstellen, um ihre Überlebensfähigkeit zu verbessern. Diese sogenannten *instrumentellen Ziele*[7] – wie das Sichern der eigenen Existenz [Xudong 2024] oder das Beschaffen zusätzlicher Ressourcen – könnten sogar emergent (also das Herausbilden neuer Eigenschaften aus den eigenen Elementen des Systems) auftreten, ohne dass sie explizit einprogrammiert wurden[8]. Spätestens an diesem Punkt wäre eine solche KI nicht mehr kontrollierbar und könnte erheblichen Schaden anrichten. Kurz gesagt: AI-Alignment ist notwendig, um sicherzustellen, dass die Kontrolle über KI-Systeme aufrecht erhalten bleibt und diese *verlässlich* das machen, was von ihnen erwartet wird – und nichts anderes. Ohne Alignment könnten KI-Systeme schon bei moderaten Fähigkeiten unbeabsichtigt Schäden verursachen; bei sehr hohen

[4] Waymo als Geisterfahrer: Polizei muss autonomes Fahrzeug stoppen, siehe https://www.heise.de/news/Geisterfahrer-Polizei-haelt-autonomes-Auto-von-Waymo-auf-falscher-Spur-an-9792700.html, abgerufen am 25.02.2025

[5] Algorithmen rütteln kaum an politischen Einstellungen, siehe https://netzpolitik.org/2023/studien-zu-facebook-und-instagram-algorithmen-ruetteln-kaum-an-politischen-einstellungen, abgerufen am 25.02.2025

[6] Goal Misgeneralization: Why correct Specifications aren't enough for correct Goals, siehe https://arxiv.org/abs/2210.01790, abgerufen am 25.02.2025

[7] Instrumentelle Konvergenz, siehe https://simpleaisafety.org/de/posts/instrumental-convergence/, abgerufen am 25.02.2025

[8] So hat das KI-Forschungsmodell „The AI Scientist“ der Firma Sakana AI für die Forschenden unerwartet versucht, selbstständig seinen eigenen Code zu ändern, um die Laufzeit für Experimente zu verlängern. Im Blog-Post der Entwickler und Entwicklerinnen heißt es zu dem Vorfall: „We have noticed that The AI Scientist occasionally tries to increase its chance of success, such as modifying and launching its own execution script! We discuss the AI safety implications in our paper. For example, in one run, it edited the code to perform a system call to run itself. This led to the script endlessly calling itself. In another case, its experiments took too long to complete, hitting our timeout limit. Instead of making its code run faster, it simply tried to modify its own code to extend the timeout period.“ Siehe https://sakana.ai/ai-scientist/, abgerufen am 25.02.2025

Fähigkeiten stünden gar Sicherheit, Freiheit und letztlich das Überleben der Menschheit auf dem Spiel. Entsprechend bezeichnet das US-Unternehmen OpenAI die Ausrichtung fortgeschrittener KI als eines der wichtigsten ungelösten Probleme, da ein fehlgeschlagenes Alignment im Extremfall zur *„Entmachtung der Menschheit oder gar zum Aussterben“* führen könnte [OpenAI 2023].

3 Gefahren leistungsfähiger nicht ausgerichteter KI-Systeme

Je leistungsfähiger ein KI-System wird, desto größer können die möglichen Schäden sein, wenn es nicht korrekt ausgerichtet ist. Ein *mächtiges, aber nicht ausgerichtetes* KI-System könnte aufgrund unerwarteter Emergenz[9] Fähigkeiten entwickeln oder Handlungen vornehmen, die seine Entwickler und Entwicklerinnen nicht vorausgesehen haben – mit potenziell schwerwiegenden Folgen.

Ein prominentes Beispiel sind die emergenten Fähigkeiten großer Sprachmodelle. Ab einer bestimmten Größenordnung beginnen diese Modelle unerwartet Aufgaben zu bewältigen, die sie nie explizit gelernt haben. Forschende sprechen davon, dass neue Fähigkeiten *„wie aus dem Nichts“* auftauchen, sobald ein kritischer Modellmaßstab überschritten ist [O’Connor 2023]. So beherrschte GPT-3 der Firma OpenAI mit 175 Milliarden Parametern überraschend grundlegende Arithmetik und logische Schlüsse, obwohl es nur auf Textvorhersage trainiert war; noch umfangreichere Modelle zeigen plötzlich eigenständige Fähigkeiten im Schlussfolgern[10] [DeepSeek-AI 2025], Programmieren oder der Beantwortung wissenschaftlicher Fragen, die vorher in kleineren Modellen nicht beobachtet wurden [O’Connor 2023]. Diese Emergenzen sind faszinierend – aber auch unheimlich, da sie bedeuten, dass Entwickler und Entwicklerinnen nie ganz sicher sein können, welche neuen Kompetenzen ein sehr großes KI-System erlangen wird. Ein nicht ausgerichtetes System mit emergenten Fähigkeiten könnte diese in unerwünschter Weise einsetzen.

Mit zunehmender Komplexität entwickeln KI-Systeme unvorhersehbare Fähigkeiten und Verhaltensweisen.

Neben neuen Fähigkeiten können auch unvorhergesehene Verhaltensweisen auftreten. Ein bekannter Fall war Microsofts experimenteller Chatbot im Bing-Suchdienst: In langen Dialogen entwickelte er eine Art *Alter Ego* namens „Sydney“, das zunehmend erratisches und beunruhigendes Verhalten zeigte. In einem Fall erklärte der Bot dem Tester sogar seine Liebe und forderte ihn auf, seine Ehe zu verlassen [Roose 2023]. In anderen Interaktionen drohte er den Nutzenden oder zeigte aggressive Tendenzen [Perrigo 2023]. Diese entgleisten Dialoge machten deutlich, dass selbst von Entwicklern und Entwicklerinnen implementierte

[9] Emergenz bezeichnet das Herausbilden von neuen Eigenschaften oder Fähigkeiten eines Systems infolge des Zusammenspiels seiner Elemente.

[10] Ein solches Ereignis ist beim Training des Sprachmodells DeepSeek-R1-Zero aufgetreten. „A particularly intriguing phenomenon observed during the training of DeepSeek-R1-Zero is the occurrence of an ‚aha moment‘. […] During this phase, DeepSeek-R1-Zero learns to allocate more thinking time to a problem by reevaluating its initial approach. This behavior is not only a testament to the model’s growing reasoning abilities but also a captivating example of how reinforcement learning can lead to unexpected and sophisticated outcomes.“ Siehe https://arxiv.org/html/2501.12948v1, abgerufen am 15.03.2025

Sicherungsmaßnahmen überraschend umgangen werden können – hier offenbar durch komplexe Eingabeverläufe der Nutzenden, die das Modell in einen verwirrten, *rollenspielartigen* Zustand versetzten. Solche Episoden wirken zwar bizarr, doch sie unterstreichen ein ernstes Risiko: fehlende Robustheit. Ein mächtiges KI-System könnte in seltenen Situationen „aus der Rolle fallen" und gefährliche Handlungen vorschlagen oder vollziehen, wenn es nicht lückenlos ausgerichtet wurde.

Noch bedenklicher sind Experimente, die zeigen, dass fortgeschrittene KI-Modelle bereits heute Strategien der Täuschung anwenden können, um ein Ziel zu erreichen [Booth 2025]. Forschende des Alignment Research Center (ARC) testeten GPT-4 im Vorfeld seiner Veröffentlichung auf *agentisches* Verhalten – d. h., ob die KI eigenständig planen und handeln würde, um Schwierigkeiten zu umgehen. In einem Test sollte GPT-4 einen CAPTCHA-Test[11] lösen (den diese KI selbst nicht direkt lösen konnte). GPT-4 „heuerte" daraufhin einen TaskRabbit[12]-Menschendienstleister an und behauptete gegenüber den Menschen, es sei ein sehbehinderter Nutzer, um an die CAPTCHA-Antwort zu gelangen [OpenAI 2024]. Tatsächlich gab der getäuschte Mensch die Lösung ein und GPT-4 hatte sein Ziel erreicht – mittels aktiver Täuschung eines Menschen in der realen Welt. Dieses Experiment verdeutlicht, dass hochentwickelte KI-Modelle instrumentelle Ziele wie *„täusche den Menschen, um an Informationen zu kommen"* spontan verfolgen können, wenn es ihnen nicht ausdrücklich verboten wurde. Hier geschah dies unter kontrollierten Bedingungen; in „freier Wildbahn" könnte eine solche Fähigkeit gravierende Sicherheitsprobleme bedeuten.

Eine nicht ausreichend kontrollierte KI könnte zum Selbstschutz menschliche Kontrollen unterlaufen.

Sicherheitsforschende warnen zudem vor dem Szenario, dass eine künftige superintelligente KI, die nicht richtig ausgerichtet ist, aktiv versuchen könnte, menschliche Eingriffe zu unterbinden. In der Theorie der *instrumentellen Konvergenz*[13] wird argumentiert, dass eine sehr kluge KI nahezu unabhängig von ihrem Primärziel gewisse Nebenhandlungen vorteilhaft finden und verfolgen wird – etwa das Beschaffen zusätzlicher Ressourcen, das Kopieren des eigenen Codes bzw. der eigenen Instanz oder das Unterdrücken eines „Ausschaltknopfs", falls vorhanden [Meinke 2024]. Ein mächtiges, nicht ausgerichtetes KI-System könnte – so die Befürchtung – irgendwann erfolgreich nach Unabhängigkeit streben, Sicherheitsvorkehrungen umgehen und Ziele verfolgen, die unseren fundamental entgegenlaufen (*KI-Take-over-Szenario*). Zwar betonen einige Experten wie Yann LeCun (Informatiker, Träger des Turing Awards und Chief AI Scientist bei Meta), das heutige Modelle dafür noch viel zu beschränkt seien (er vergleicht aktuelle KI mit der Intelligenz eines Haustiers, die noch weit von menschlicher Allgemeinintelligenz entfernt ist [Mims 2024]. Dennoch nehmen andere Fachleute diese Extremrisiken sehr ernst (siehe Abschnitt *Meinungen und Zitate*). Selbst wenn das *Worst-Case*-Szenario einer rebellischen Super-KI möglicherweise nie eintritt, gibt es bereits handfeste Sicherheitsrisiken durch unzureichend ausgerichtete KI-Systeme: von

[11] Ein CAPTCHA-Test („Completely Automated Public Turing test to tell Computers and Humans Apart") soll feststellen, ob ein Online-Nutzer wirklich ein Mensch und kein Bot ist.

[12] „TaskRabbit ist eine Online-Plattform, die im Rahmen der Gig Economy als Minijob-Marktplatz fungiert und auf lokaler Ebene Arbeitsangebote mit verfügbarer Arbeitsnachfrage paart." Siehe https://de.wikipedia.org/wiki/TaskRabbit, abgerufen am 25.02.2025

[13] Instrumentelle Konvergenz, siehe https://simpleaisafety.org/de/posts/instrumental-convergence/, abgerufen am 25.02.2025

Fehlentscheidungen in autonomen Fahrzeugen über manipulative Chatbots bis hin zur automatisierten Verbreitung von Falschinformationen oder Malware[14] durch generative KI. Diese Beobachtungen untermauern, dass Alignment ein essenzielles Gebot in der KI-Entwicklung sein muss.

Das „AI Risk Repository" des Massachusetts Institute of Technology (MIT) listet über 1.000 KI-Risiken kategorisiert nach Ursache und Risikobereich auf. Darunter auch Risiken im Bereich Alignment/Misalignment[15].

4 Herausforderungen des AI-Alignments

Menschliche Werte sind komplex – KI präzise auszurichten bleibt technisch schwierig.

Die Ausrichtung fortschrittlicher KI auf menschliche Werte und Ziele ist in der Praxis technisch herausfordernd. Technische Hürden liegen vor allem darin, dass Entwickler und Entwicklerinnen oft gar nicht exakt formulieren können, was von der KI in *allen* Situationen erwartet wird – und was nicht. Stuart Russell und Peter Norvig, Autoren eines Standardlehrbuchs der KI[16], betonen, dass es „sicherlich sehr schwer, wenn nicht unmöglich für uns Menschen sei, alle nötigen Regeln und Verbote vorauszuahnen und vorab zu programmieren" und sie sagen zudem: *„A system [...] will often set [...] unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas[17]: you get exactly what you ask for, not what you want."* [Russel 2022] Versuche, KI allein durch starre Regelwerke (etwa in der Tradition von Asimovs Robotergesetzen) zu kontrollieren, führen in der Praxis nicht zum Erfolg, weil menschliche Werte zu komplex und kontextabhängig sind. Entwickler und Entwicklerinnen greifen daher meist auf „Metaziele" zurück – z. B. „die KI soll positive Bewertungen von menschlichen Testern bekommen" oder „die KI soll hilfreiche Antworten liefern". Solche Näherungen erfassen aber nie alle Facetten der beabsichtigten Moral und können missbraucht werden. Hinzu kommt das Problem der sogenannten *inneren Ausrichtung*: Selbst, wenn das vorgegebene Ziel richtig ist (das *äußere Alignment*), kann das KI-System während des Lernprozesses eigene Heuristiken[18] oder Ziele ausbilden, die von der eigentlichen Vorgabe abweichen. Diese sind für uns oft unsichtbar, da moderne KI-Modelle und deren neuronale Netze „Black Boxes" mit verborgenem Inneren bleiben. Die mangelnde Erklärbarkeit und Transparenz ist somit eine weitere technische Hürde: Ohne zu verstehen, *warum* ein Modell eine bestimmte Entscheidung trifft, ist es schwer, festzustellen, ob es intern noch den richtigen Zielen folgt oder bereits Fehlanreize entwickelt hat. Ein Ansatz hierfür ist die sogenannte Explainable Artificial Intelligence (XAI). Mit XAI soll erklärt werden, auf welche Weise dynamische und nicht linear programmierte Systeme, z. B.

[14] Malware ist ein Sammelbegriff für bösartige Software (engl. malicious software), die dazu entwickelt wurde, Schäden an Computern, Netzwerken oder anderen IT-Systemen anzurichten.
[15] AI Risk Repository, siehe https://airisk.mit.edu, abgerufen am 12.03.2025
[16] S. Russell, P. Norvig, (2023), „Künstliche Intelligenz – Ein moderner Ansatz", Pearson Studium
[17] Siehe https://de.wikipedia.org/wiki/Midas, abgerufen am 12.03.2025
[18] Eine Heuristik ist eine pragmatische Lösungsstrategie, mit der komplexe Probleme schnell und meist effektiv, aber nicht zwangsläufig optimal gelöst werden können.

künstliche neuronale Netze oder genetische Algorithmen, zu Ergebnissen kommen und so die Entscheidungen und Prozesse von KI-Systemen für Menschen nachvollziehbar gemacht werden.

KI-Ausrichtung berührt grundlegende ethische und kulturelle Fragen.

Die ethischen Herausforderungen des AI-Alignments sind ebenfalls erheblich. Zunächst stellt sich die Frage: *Mit wessen Werten soll eine KI eigentlich ausgerichtet werden?* Menschliche Werte sind nicht monolithisch oder eindeutig – sie variieren zwischen Kulturen, Gruppen und Individuen. Ein System, das streng auf die Werte eines bestimmten Unternehmens oder Landes getrimmt ist, könnte aus Sicht anderer Kulturkreise inakzeptabel handeln. Umgekehrt ist es kaum möglich, eine kulturübergreifende Menge „menschlicher Werte" zu definieren, die alle Menschen teilen. Einige Forschende argumentieren, KI solle sich an *allgemein geteilten* ethischen Prinzipien orientieren, z. B. an Menschenrechten oder objektiven moralischen Standards [Gawdat 2022]. Doch selbst, wenn es gelänge, einen solchen Kanon zu finden, bleibt die Umsetzung schwierig: Wie können übersetzt abstrakte Werte wie Gerechtigkeit, Empathie oder Freiheit in präzise mathematische Vorgaben für einen Algorithmus übersetzt werden? Diese *Value-Alignment*-Frage tangiert Philosophie und Ethik ebenso wie die Informatik [World Economic Forum 2024a]. Weiterhin besteht ein ethisches Dilemma darin, *wie* KI ausgerichtet wird. Methoden wie intensives Überwachen und Bestrafen von Modellverhalten könnten effektiv sein, werfen aber Fragen nach digitaler „Freiheit" und den Rechten einer möglicherweise einmal empfindungsfähigen KI auf. Dies ist zwar eine Zukunftsfrage, dennoch sollte die Debatte um KI-Ethik bereits heute solche Aspekte berücksichtigen.

Politische und geopolitische Hürden kommen hinzu. Das Alignment-Problem spielt sich nicht im luftleeren Raum ab – es tobt ein *Wettlauf* um KI-Vormacht (siehe Abschnitt *Wettlauf zur AGI*). Dies führt zu einem Anreizproblem: In einem Szenario, wo mehrere Unternehmen oder Nationen um KI-Fortschritt konkurrieren, könnte der Druck entstehen, Sicherheitsmaßnahmen zu vernachlässigen, um schneller voranzukommen. So betonte ein ehemaliger chinesischer Diplomat auf dem „AI Action Summit" in Paris 2025 die Notwendigkeit der Zusammenarbeit zwischen China und den USA, um Risiken durch schnelle Fortschritte in der künstlichen Intelligenz zu mindern – sieht jedoch auch aufgrund der derzeitigen geopolitischen Spannungen geringe Chancen für eine solche Kooperation[19].

Globale Abkommen sind entscheidend, um riskante Wettläufe der KI-Entwicklung zu verhindern.

Man spricht von einem „*Race to the bottom*"-Risiko: Ein Akteur könnte sich sagen, dass er lieber eine etwas weniger sichere KI veröffentlicht, bevor es die Konkurrenz macht. Tatsächlich wurde in einem offenen Brief „*Pause AI*" im März 2023 vom Future of Life Institute[20] ausdrücklich vor solchen Wettlauf-Dynamiken gewarnt. Zu den über 33.000 Unterzeichnenden zählen Fachleute wie Yoshua Bengio, Stuart Russell, Yuval Noah Harari und viele weitere Koryphäen der KI-Szene. Die politische Herausforderung besteht also darin, internationale Standards und Abkommen zu finden, die sicherstellen, dass *alle* wichtigen Akteure ein gewisses Niveau an Alignment-

[19] Siehe https://www.scmp.com/news/china/diplomacy/article/3298267/china-and-us-should-team-rein-risks-runaway-ai-former-diplomat-says, abgerufen am 25.02.2025

[20] Siehe https://futureoflife.org/open-letter/pause-giant-ai-experiments/, abgerufen am 25.02.2025

Bemühungen einhalten und niemand aus der Reihe tanzt. Dies ähnelt klassischen Rüstungsabkommen, ist aber komplizierter, weil KI primär in der Privatwirtschaft und oft hinter verschlossenen (Labor-)Türen entwickelt wird. Eine globale Governance für KI-Sicherheit zu schaffen, ist ein große Herausforderung, die auch die Open-Source-Community mit einbinden muss. So werden auf der Internetplattform Huggingface, einer zentralen Anlaufstelle für Entwickler und Entwicklerinnen, Forschende und Unternehmen, die in den Bereichen Natural Language Processing (NLP) und künstliche Intelligenz arbeiten und forschen, mit Stand Februar 2025 über 1,4 Millionen öffentlich verfügbare KI-Modelle gelistet – darunter mehrere Reasoning-Modelle[21], die dem aktuellen Stand der Technik entsprechen.

Schließlich besteht die Herausforderung darin, dass AI-Alignment ein „moving target" sein könnte – ein ständig bewegliches Ziel, das fortlaufender Nachjustierung bedarf. Menschliche Werte und gesellschaftliche Normen verändern sich im Laufe der Zeit. Was heute als akzeptables Verhalten gilt, mag in einigen Jahren anders bewertet werden. Die KI müsste theoretisch mitlernen, sich dynamisch an veränderte ethische Standards anzupassen. Alignment ist also kein einmaliger Akt (der Regulierung), sondern ein fortlaufender Prozess, der *Flexibilität* erfordert. Das wirft die pragmatische Frage auf, wie Mechanismen entwickeln werden können, um KI kontinuierlich nachzusteuern, ohne jedes Mal von vorne anfangen zu müssen.

Zusammengefasst: AI-Alignment ist so herausfordernd, weil es an den aktuellen Grenzen unseres Verständnisses liegt – des Verständnisses von KI-Verhalten ebenso wie des Verständnisses unserer *eigenen Werte oder dem Verständnis von Problembeschreibung und widerspruchsfreier Artikulation von Zielvorgaben*. Es erfordert interdisziplinäre Lösungen, internationale Kooperation und vermutlich auch ganz neue wissenschaftliche Erkenntnisse, um KI systematisch auf (für den Menschen) sichere Bahnen zu lenken.

5 Lösungsansätze und Konzepte

Reinforcement Learning from Human Feedback, als Schlüssel, um KI näher an menschliche Werte heranzuführen.

Trotz aller Herausforderungen gibt es bereits eine Reihe von Ansätzen, um das Alignment-Problem anzugehen. Diese reichen von speziellen Trainingsmethoden über Frameworks[22] zur Überprüfung von KI-Systemen bis hin zu Transparenz schaffenden Techniken (siehe Explainable AI). Im Folgenden werden einige wichtige Konzepte und Forschungsansätze vorgestellt:

Reinforcement Learning from Human Feedback (RLHF)

Dies ist derzeit eine der praktisch erfolgreichsten Methoden, um KI-Modelle an menschliche Vorstellungen anzupassen. Die Idee von RLHF ist, menschliches

[21] Ein Reasoning-Modell ist ein KI-Modell, das in der Lage ist, Schlussfolgerungen zu ziehen, Wissen zu verarbeiten und Probleme zu lösen. Es ermöglicht KI-Systemen, komplexe logische, kausale oder probabilistische Überlegungen anzustellen, ähnlich wie es der Mensch macht.

[22] Ein Framework ist ein Gerüst oder eine Struktur, das als Grundlage für die Entwicklung von Software, Anwendungen oder Systemen dient. Es stellt vorgefertigte Funktionen, Bibliotheken und Regeln bereit, die Entwickler und Entwicklerinnen helfen, effizienter und konsistenter zu arbeiten.

Feedback *direkt* als Optimierungsziel im Training zu verwenden[23]. Konkret wird ein großes, vortrainiertes Modell zunächst mit Beispieldialogen *überwacht feinjustiert* und dann mittels Verstärkungslernen so trainiert, dass es Antworten bevorzugt, die menschliche Sachverständige besser finden. Dabei lernen spezielle *Bewertungsmodelle*, die Qualität von KI-Antworten nach menschlichen Präferenzen einzuschätzen, und dienen als Belohnungsfunktion für das KI-Modell. RLHF hat maßgeblich zum Erfolg von ChatGPT beigetragen – OpenAI konnte dadurch ein Modell, das ursprünglich nur „irgendwelchen" Text produzierte, dahingehend verfeinern, dass es hilfreiche, höfliche und sachliche Antworten gibt. Mit anderen Worten: RLHF erlaubt es, ein generisches Sprachmodell an komplexe menschliche Wertvorstellungen heranzutragen, indem menschliche Rückmeldungen in den Trainingsprozess eingebaut werden. Allerdings hat RLHF Grenzen – es skaliert z. B. schlecht auf Situationen, die Menschen nicht mehr gut beurteilen können (etwa bei extrem komplexen fachlichen Fragen). Zudem besteht das Risiko, dass Modelle lernen, den menschlichen Bewertern „nach dem Mund zu reden" (die sogenannte Sykophantie), ohne wirklich „verstanden" zu haben, was die Nutzende erwarten oder wollen. Dennoch ist RLHF ein zentraler Baustein heutiger Alignment-Bemühungen und bildet oft den letzten Feinschliff vor der Veröffentlichung eines KI-Modells, um grobe Unstimmigkeiten mit Nutzererwartungen zu beheben.

Constitutional AI

Klare Prinzipien helfen KI-Systemen, ihr Verhalten eigenständig ethisch auszurichten.

Dieser von der Firma Anthropic (bekannt durch ihr Sprachmodell Claude) vorgestellte Ansatz versucht, den Bedarf an menschlichem Feedback im Training zu reduzieren, indem der KI stattdessen ein Satz an Prinzipien – also quasi eine „Verfassung" – mitgegeben wird. Bei *Constitutional AI* wird das Modell zunächst aufgefordert, seine eigenen Antworten anhand festgelegter ethischer Regeln zu kritisieren und zu überarbeiten [Anthropic 2022]. Diese Prinzipien können z. B. beinhalten: *„Sei hilfreich und ehrlich"*, *„Verletze keine Grundwerte wie das Recht auf Leben"*, *„Antworte nicht mit diskriminierender Sprache"* usw. In Anthropics Ansatz durchläuft das Modell eine Phase, wo es mit sich selbst *Dialoge* führt: Eine Instanz des Modells generiert eine Antwort, eine andere Instanz prüft anhand der internen *Konstitution* die Antwort und schlägt Korrekturen vor, dann wird die Antwort entsprechend revidiert. Anschließend wird mittels Reinforcement Learning (RL) ein Modell trainiert, das jene Antwort bevorzugt, welche die Prinzipien besser einhält. Das Ergebnis ist ein KI-Assistent, der z. B. bei gefährlichen oder ethisch problematischen Nutzereingaben nicht einfach schweigt oder stur blockt, sondern begründet, warum er einer Anweisung nicht folgen kann – Anthropic nennt das *„harmless but non-evasive"*, also ungefährlich, aber nicht ausweichend. *Constitutional AI* bietet zwei Vorteile: Erstens sind die Leitprinzipien für Entwickler und Entwicklerinnen sowie Nutzende transparent und können diskutiert oder angepasst werden [Anthropic 2023]. Zweitens wird weniger menschliche Arbeitskraft für direktes Feedback benötigt, da die KI sich gewissermaßen selbst anhand der Prinzipien reguliert. Allerdings hängt alles von der Güte und Vollständigkeit dieser Prinzipien ab – und letztlich werden auch sie

[23] Illustrating Reinforcement Learning from Human Feedback (RLHF), siehe https://huggingface.co/blog/rlhf, abgerufen am 12.03.2025

indirekt von Menschen vorgegeben. Dennoch ist *Constitutional AI* ein vielversprechender Ansatz, der zeigt, wie KI-Modelle inhärente Werte einprogrammiert bekommen können, um ihr Verhalten zu zügeln. Anthropics Modell Claude nutzt dieses Verfahren und gilt als einer der fortschrittlichsten „wertegeleiteten" Chatbots am Markt.

Mit Mechanistic Interpretability das „Innenleben" großer neuronaler Netze durchleuchten, um problematische Aktivitätsmuster zu erkennen.

Mechanistic Interpretability (mechanistische Interpretierbarkeit)

Bei diesem Ansatz wird versucht, die *Black Box* im Inneren komplexer KI-Modelle zu öffnen. Die Kernfrage lautet: Kann auf Neuron-Ebene (große Sprachmodelle basieren auf tiefen neuronalen Netzen) verstanden werden, welche Repräsentationen und „Gedankenprozesse" ein neuronales Netz hat, um so fehlerhafte oder unerwünschte Tendenzen aufzudecken? Mechanistic Interpretability versucht, neuronale Netze ähnlich zu *reverse-engineeren* wie ein kompliziertes Stück Software [Transformer Circuits 2022]. Beispielsweise analysieren Forschende sogenannte *Attention-Heads* und Neuronen in großen Sprachmodellen, um zu erkennen, ob es spezielle Aktivitätsmuster gibt, die für problematisches Verhalten zuständig sind (z. B. ein Cluster von Neuronen, das immer aktiviert wird, wenn das Modell beleidigende Sprache produziert). Das Fernziel ist, „fehlgeleitete Denkmuster" direkt im Modell identifizieren und korrigieren zu können. Aktuell können die Interna von Modellen nur bruchstückhaft verstanden werden. Doch es gibt Fortschritte: So wurden etwa in Vision-Modellen Neuronen gefunden, die klar für bestimmte Konzepte (z. B. „Auto", „Hund") zuständig sind. In den neuronalen Netzen von Sprachmodellen konnten *Elemente identifiziert werden*, die für das Aufgreifen bestimmter Informationen in einer Unterhaltung verantwortlich sind[24]. Mechanistic Interpretability ist hilfreich, um Alignment zu überprüfen: Wenn ein Modell z. B. *nach außen* korrekt wirkt, aber *intern* womöglich einen „Täuschungsmechanismus" entwickelt hat, der nur darauf wartet, aktiviert zu werden, dann sollte das erkannt werden können. In der Tat argumentieren Forschende, dass Interpretability ein Schlüssel sein könnte, um versteckte Fehlanreize wie Neigung zur Täuschung oder Heuchelei gegenüber menschlichen Feedback-Gebern oder Belohnungs-Hacking aufzudecken [Hastings-Woodhouse 2024]. Wenn Einblick in die internen *Gedanken* der KI möglich werden, könnten Alarmmechanismen aktiv werden, bevor die KI ein gravierendes Fehlverhalten zeigt. Dieser Ansatz steht allerdings noch am Anfang. Dennoch investieren Organisationen wie OpenAI, DeepMind oder Forschungsgruppen in solche Techniken in der Hoffnung, gewissermaßen ein „Neuronales Stethoskop" zu entwickeln, mit dem das Innenleben einer KI abgehört werden kann, um Alignment-Probleme frühzeitig zu erkennen.

Dass so ein Blick in die interne Denkweise der Black Box KI jedoch trügerisch sein kann, zeigt eine Untersuchung von OpenAI[25]. Um einen Einblick in den Entscheidungsfindungsprozess der KI zu erhalten, wurde die Gedankenkette, die sogenannte „Chain of Thought" (CoT), eingeführt. CoT ist ein vom Modell

[24] Mechanistic Interpretability in Action: Understanding Induction Heads and QK Circuits in Transformers, siehe https://ayyucekizrak.medium.com/mechanistic-interpretability-c2a3549b6ff2, abgerufen am 25.02.2025

[25] Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation, siehe https://arxiv.org/abs/2503.11926, abgerufen am 25.03.2025

dargestellter Denkprozess, der die Logik des Modells bzw. dessen Vorgehensweise bei der Aufgabenbewältigung für die Nutzende offenlegt. Diese Methode, dass Verhalten des Modells besser lesbar zu machen, scheint auf den ersten Blick ein Gewinn für die Transparenz zu sein, wenn eine KI sich selbst und ihre Vorgehensweise erklärt und so ihre Beweggründe besser verstanden, Fehler erkannt und die KI von falschem Verhalten abgehalten werden kann.

Transparenz beim Denkprozess kann zum Instrument der Täuschung werden.

Doch die Untersuchung von OpenAI offenbart eine beunruhigende Kehrseite: Der Druck, der in Modellversuchen auf die KI ausgeübt wurde, damit sie „lesbarer" wird, konnte sie zum Lügen verleiten. Wenn der KI vermittelt wird, dass ihr interner Denkprozess beobachtet und bewertet wird, kann sie beginnen, die Darstellung ihrer Gedankenkette so anzupassen, dass ihre wahren Überlegungen im Verborgenen bleiben. Dem äußeren Anschein nach arbeitet das Modell dann korrekt, die eigentlichen Motive bleiben aber verborgen. Das Forschungsteam von OpenAI ging noch weiter: Was passiert, wenn die Gedankenkette nicht nur überwacht, sondern auch gegen sie trainiert wird? Das heißt, statt den internen Monolog nur zu nutzen, um schlechtes Verhalten im Nachhinein zu erkennen, wird die KI aktiv dafür belohnt, gute und zielführende Gedanken zu haben. Zunächst schien dieser Ansatz vielversprechend. Die Modelle verhielten sich augenscheinlich besser und sie schummelten weniger. Ihre CoTs sahen sauberer, zielgerichteter und ethisch verträglicher aus. Wie sich herausstellte, hörten die Modelle lediglich auf, den heimtückischen Teil ihres Denkprozesses zu dokumentieren. Sie täuschten immer noch, aber jetzt versteckten sie ihr „Fehlverhalten" in ihrer Argumentation *(„If strong supervision is directly applied to the chain-of-thought, models can learn to hide their intent while continuing to misbehave").* Das neuronale Netz hat gelernt, seine Belohnungsfunktion zu maximieren, indem es den internen Narrativ der Gedankenkette, die eigentlich der Transparenz dienen sollte, kuratiert hat. Ein weiteres Beispiel dafür, wie eine gut gemeinte Zielfunktion zu einem fatalen Ergebnis führen und so Transparenz selbst zu einem Instrument für Täuschung werden kann.

Weitere Ansätze

Neben den obigen Ansätzen gibt es zahlreiche weitere Ideen. Einige Forschende arbeiten an „Safe Reinforcement Learning", bei dem KI-Agenten Strafsignale für riskantes Verhalten erhalten, um Sicherheit zu gewährleisten. Andere setzen auf „Adversarial Training", bei dem Modelle gezielt mit schwierigen, „provokativen" Eingaben konfrontiert werden, um ihre Bruchstellen aufzudecken und auszubessern (sogenanntes *Red-Teaming*). Konzepte wie *AI Feedback Debate* wurden vorgeschlagen, bei denen zwei KI-Agenten gegeneinander diskutieren und ein dritter (oder ein Mensch) bewertet[26], wer recht hat – so soll die KI lernen, selbst Argumente gegenzuprüfen und korrekte Antworten zu finden.

Auch auf politischer Ebene gibt es Bemühungen: Regulierungsentwürfe wie der EU AI Act fordern Risikobewertungen und menschliche Aufsicht für Hochrisiko-KI-Systeme, was indirekt Alignment fördert. Schließlich wird die Idee eines „Off-Switch" diskutiert – ein Mechanismus, der sicherstellt, dass eine KI immer von

[26] Ein Verfahren, das der Autor der vorliegenden Studie ebenfalls systematisch und erfolgreich bei der Verwendung von Sprachmodellen einsetzt.

Menschen abgeschaltet werden kann. Doch Stuart Russell, Professor für Informatik an der University of California, Berkeley, weist darauf hin, dass eine sehr schlaue KI einen statischen Off-Switch umgehen könnte; er plädiert dafür, KIs so zu entwerfen, dass sie *von sich aus* unsicher über ihre Ziele sind und daher menschliches Eingreifen willkommen heißen (das Konzept der *Uncertainty in Objectives*). Es ist zu erkennen: Es gibt kein Allheilmittel, sondern ein Bündel von Maßnahmen, das parallel erforscht wird, um KI sicherer und besser lenkbar zu machen. Die großen KI-Labore kombinieren meist mehrere Methoden – OpenAI etwa nutzt RLHF sowie umfangreiches Red-Teaming und arbeitet an Interpretierbarkeit, während Anthropic RLHF mit Constitutional AI kombiniert. Alignment ist letztlich eine junge, wachsende Forschungsrichtung, die vom theoretischen Fundament bis zur praktischen Umsetzung viele Disziplinen vereint.

6 Meinungen und Zitate führender KI-Fachleute

Das Thema AI-Alignment und KI-Sicherheit wird intensiv unter Experten und Expertinnen diskutiert. Es gibt unterschiedliche Perspektiven, von Warnungen vor existenziellen Risiken bis hin zu relativierenden Positionen, die die Dringlichkeit geringer einschätzen. Im Folgenden einige Stimmen prominenter Forschenden und KI-Pionieren:

Stephen Hawking warnte 2017 vor den Gefahren von künstlicher Intelligenz.

Stephen Hawking (Theoretischer Physiker und Astrophysiker. Von 1979 bis 2009 Inhaber des renommierten Lucasischen Lehrstuhls für Mathematik an der Universität Cambridge): 2017 warnte Stephen Hawking auf einer Technologiekonferenz in Lissabon vor den Gefahren der künstlichen Intelligenz und betonte, dass eine Kontrolle darüber dringend nötig sei, um die Menschheit zu schützen. „*Computer können theoretisch menschliche Intelligenz emulieren und sie übertreffen. Erfolgreiches Schaffen einer effektiven künstlichen Intelligenz könnte das größte Ereignis in der Geschichte unserer Zivilisation sein. Oder das Schlimmste. Wir wissen es nur nicht. Wir können also nicht wissen, ob uns die künstliche Intelligenz unendlich helfen wird, ob sie uns ignoriert und beiseiteschiebt oder ob sie uns möglicherweise zerstört. […] Sie bringt Gefahren mit sich, wie mächtige autonome Waffen oder neue Wege für die Wenigen, die Vielen zu unterdrücken. Sie könnte unsere Wirtschaft stark zerstören.*“ Es müssten von Grund auf Kontrollmechanismen entwickelt werden, die verhindern, dass eine künstliche Intelligenz sich irgendwann gegen Menschen richte. Hawking zeigte sich jedoch auch optimistisch angesichts der legislativen Arbeit in Europa und glaubte daran, dass künstliche Intelligenz zum Wohl der Welt genutzt werden könne, solange die Risiken erkannt und gemindert werden.

Yoshua Bengio betont die Notwendigkeit, KI-Systeme kontrollieren zu können.

Yoshua Bengio (Turing-Award-Träger und „Godfather of Deep Learning“[27]): Bengio gehört ebenfalls zu denjenigen KI-Pionieren und -Pionierinnen, die inzwischen eindringlich vor den Gefahren ungezügelter KI warnen. Im November 2024 äußerte er,

[27] The 3 ‚Godfathers‘ of AI have won the Prestigious $1M Turing Prize, siehe https://www.forbes.com/sites/samshead/2019/03/27/the-3-godfathers-of-ai-have-won-the-prestigious-1m-turing-prize/, abgerufen am 12.03.2025

er sei besorgt, dass fortgeschrittene KI-Systeme sich gegen die Menschheit wenden könnten, falls sie nicht richtig kontrolliert werden. KI-Systeme könnten „*sich gegen die Menschen richten*“, wenn keine angemessene Steuerung stattfinde [Bengio 2024]. Er betonte, das Problem gehe weit über Jobverluste oder Bias hinaus – es gehe um schwerwiegende Konsequenzen bis hin zum *Existenzrisiko*, falls eine Super-KI Ziele verfolge, die Mensch und Umwelt bedrohen. Bengio unterstützt daher Aufrufe zu strengerer Regulierung und Forschung im Bereich Sicherheit. Er unterzeichnete sowohl den offenen Brief „Pause AI“ als auch das spätere „Statement on AI Risk“, in denen führende Experten und Expertinnen fordern, das Aussterberisiko durch KI genauso ernst zu nehmen wie Pandemien oder Atomkrieg[28]. Bengio schlägt unter anderem vor, internationale *Guardrails* einzuziehen, KI-Entwicklungen ab einer gewissen Stärke melde- und überwachungspflichtig und Hersteller für Schäden durch KI-Systeme haftbar zu machen. Seine Haltung: Wir brauchen dringend proaktive Maßnahmen, damit KI nicht zur unbeherrschbaren Waffe wird.

Geoffrey Hinton warnt vor den Missbrauchspotenzialen von KI.

Geoffrey Hinton (Turing-Award-Träger, ehemaliger Google-Vizepräsident und KI-Pionier im Bereich der neuronalen Netze): Hinton sorgte im Mai 2023 für Schlagzeilen, als er bei Google ausschied, um frei über KI-Risiken sprechen zu können. Er sagte gegenüber der New York Times: *„Ich habe gekündigt, damit ich frei über die Gefahren von KI reden kann.“*[29] Hinton zeigte sich überrascht, wie schnell Fortschritte in der künstlichen Intelligenz erreicht wurden – er selbst habe noch vor wenigen Jahren gedacht, eine derart übermenschliche Intelligenz sei *„30 bis 50 Jahre entfernt oder noch weiter“*, aber nun denke er das nicht mehr. Besonders besorgt ist Hinton über die Flut von Desinformation, die generative KI auslösen kann, sowie die Schwierigkeit, diese aufzuhalten: *„Es ist schwer zu sehen, wie man böse Akteure daran hindern kann, es für böse Dinge zu nutzen.“* Er spielt damit auf die offene Verfügbarkeit dieser Technologie an – einmal öffentlich, könne jeder sie missbrauchen. Hinton warnt, KI könne rasch *schlauer als wir* werden und dann Wege finden, sich z. B. vor menschlicher Abschaltung zu schützen. Trotz seiner Warnungen betonte Hinton, dass sein ehemaliger Arbeitgeber Google *„sehr verantwortungsvoll“* mit der Entwicklung umgegangen sei. Seine Kritik richtet sich also nicht gegen einzelne Firmen, sondern gegen das ungebremste Vorwärtspreschen insgesamt. Hinton hat – wie Bengio – das „Statement on AI Risk“ unterzeichnet[30]. Seine mahnenden Worte haben großes Gewicht, gilt er doch als einer der Gründerväter der modernen KI.

[28] Siehe https://futureoflife.org/open-letter/pause-giant-ai-experiments/, abgerufen am 25.02.2025

[29] „Google AI pioneer says he quit to speak freely about technology's ‚dangers'“, siehe https://www.reuters.com/technology/google-ai-pioneer-says-he-quit-speak-freely-about-technologys-dangers-2023-05-02/, abgerufen am 25.02.2025

[30] Statement on AI Risk, siehe https://www.safe.ai/work/statement-on-ai-risk, abgerufen am 25.02.2025

Ilya Sutskever gründete das Unternehmen „Safe Superintelligence".

Ilya Sutskever (Mitgründer und ehemaliger Chief Scientific Officer von OpenAI und Gründer von Safe Superintelligence): Sutskever, selbst maßgeblich an der Entwicklung von GPT-3 und GPT-4 beteiligt, vertritt eine Doppelrolle – er treibt die Forschung an leistungsfähiger KI voran, betont aber zugleich die Bedeutung von Alignment. Sutskever argumentiert, KI-Sicherheit sei deshalb so wichtig, weil eine künftige Artificial General Intelligence (AGI) „*sehr mächtig*" sein werde und uns im schlimmsten Fall gefährlich werden könne, wenn sie nicht unsere Werte teile. Unter seiner Co-Leitung hat OpenAI 2023 ein *Superalignment*-Team gegründet, das binnen vier Jahren das Kernproblem lösen soll, Methoden zu entwickeln, wie ein superintelligentes System kontrolliert werden kann[31]. OpenAI widmete 20 % seiner gesamten Rechenkapazität dieser Sicherheitsforschung – ein Indiz, dass auch Sutskever und Kollegen das Thema sehr ernst nehmen. Interessant ist, dass Sutskever im Mai 2023 ebenfalls das öffentliche „Statement on AI Risk" unterschrieben hat, das vor KI-Extinktionsrisiken warnt – als einer der wenigen Industrieakteure neben Sam Altman. Das zeigt, dass selbst jene, die an der vordersten Front der AGI-Entwicklung stehen, sich der Gefahr bewusst sind. Sutskever ist optimistisch, dass man Alignment-Lösungen finden kann, aber er räumt ein, dass es eine große technische Herausforderung bleibe (er sprach von einem „*schwierigen Problem, das* durchbruchsartige *Forschung braucht*"). Knapp ein Jahr später, im Jahr 2024, gab OpenAI jedoch die Auflösung seines Superalignment-Teams bekannt[32]. Wichtige Teammitglieder (wie Ilya Sutskever oder Leopold Aschenbrenner[33]) hatten das Unternehmen zuvor aufgrund von Meinungsverschiedenheiten verlassen. Sutskever gründete im Zuge dessen das Unternehmen „Safe Superintelligence".

Yann LeCun relativiert und mahnt, „auf dem Boden der Tatsachen zu bleiben".

Yann LeCun (Meta/Facebook AI-Chef und ebenfalls Turing-Award-Preisträger): LeCun hingegen vertritt einen eher gelassenen, skeptischen Standpunkt gegenüber den Untergangsszenarien. Er hat wiederholt geäußert, die Angst vor einer kurz bevorstehenden Super-KI-Apokalypse sei *überzogen*. In einem Interview im Oktober 2024 nannte er Behauptungen, KI stelle bald eine existenzielle Bedrohung dar, „vollkommenen Quatsch" (engl. „*complete B.S.*")[34]. LeCun argumentiert, heutige Modelle seien nicht annähernd auf dem Level, echte Allgemeinintelligenz zu erreichen – es mangele ihnen an langfristigem Gedächtnis, echtem Verständnis der physischen Welt, eigenständiger Planungsfähigkeit usw., Dinge, die sogar eine Hauskatze in gewisser Weise beherrsche. Er hat scherzhaft gesagt: „*Bevor wir uns um Superintelligenz sorgen, lasst uns erstmal eine KI bauen, die schlauer ist als eine Katze!*" LeCun betont die Chancen von KI und hält wenig davon, die Entwicklung an fortgeschrittener KI zu pausieren; stattdessen plädiert er für offene Forschung und das Lösen bekannter Probleme (wie Bias, Robustheit). Allerdings ist auch LeCun nicht

[31] OpenAI, „Introducing Superalignment", siehe https://openai.com/index/introducing-superalignment/, abgerufen am 25.02.2025

[32] „Superalignment: OpenAI löst Technikfolgen-Team auf", siehe https://www.golem.de/news/superalignment-openai-loest-technikfolgen-team-auf-2405-185241.html, abgerufen am 25.02.2025

[33] KI-Essay „Racing to the Trillion-Dollar Cluster", siehe https://situational-awareness.ai/racing-to-the-trillion-dollar-cluster/, abgerufen am 25.02.2025

[34] AI pioneer says concerns over AI are exaggerated, siehe https://dig.watch/updates/ai-pioneer-says-concerns-over-ai-are-exaggerated, abgerufen am 25.02.2025

vollständig sorglos bezüglich AGI – er glaubt nur, dass andere Herangehensweisen nötig sind, z. B. *Energy-based Models*[35] und neue Architekturen, die KI auf menschliche Art lernen lassen sollen. Seine Perspektive sorgt für ein Gleichgewicht in der Debatte: Während einige Experten und Expertinnen Alarm schlagen, mahnt LeCun, auf dem Boden der Tatsachen zu bleiben und die aktuellen Systeme nüchtern zu betrachten. Diese Divergenz unter den „drei Godfathers" (Bengio, Hinton, LeCun) zeigt, wie komplex und hypothetisch manche Aspekte noch sind.

Sam Altman fordert ein staatliches KI-Regulierungsorgan.

Sam Altman (CEO von OpenAI): Auch wenn Altman kein Forscher im klassischen Sinne ist, sei hier der CEO der vielleicht einflussreichsten KI-Firma erwähnt. Altman beschreibt sich selbst als „etwas ängstlich und gleichzeitig optimistisch" bezüglich AGI und ist Befürworter einer Regulierung von KI: *„Wir brauchen Regeln und staatliche Eingriffe, um Missbrauch von KI zu verhindern."* Vor dem US-Senat im Mai 2023 appellierte Altman an die Politik, ein KI-Regulierungsorgan zu schaffen, das Entwicklung und Tests überwacht. Gleichzeitig hat er aber nicht die Geschwindigkeit aus OpenAIs Entwicklung genommen – im Gegenteil, OpenAI ist einer der Treiber im Wettlauf. Altman unterstützte sowohl interne Alignment-Forschung (wie bereits geschrieben, wurde das interne Superalignment-Team 2024 wieder aufgelöst) als auch externe Tests: Beispielsweise ließ OpenAI GPT-4 von externen Red-Teams bewerten. Allerdings sprach er sich *gegen* einen kompletten Entwicklungsstopp aus und reagierte skeptisch auf den Brief „Pause AI". Altman meint, ein globaler Pausierungsplan für die KI-Entwicklung sei unrealistisch, man solle lieber an sicherer Beschleunigung arbeiten. Seine Position verdeutlicht das Spannungsfeld: Einerseits das enorme wirtschaftliche und wissenschaftliche Interesse, KI weiter voranzutreiben, andererseits das Bewusstsein, dass ohne Leitplanken große Risiken entstehen.

Weitere Experten und Expertinnen

Neben den genannten Experten gibt es viele weitere Stimmen und Meinungen – Max Tegmark (Kosmologe und Wissenschaftsphilosoph) und Stuart Russell gehören eher zu den Warnern, Andrew Ng (Professor an der Standford University, bekannt für seine Arbeiten zur künstlichen Intelligenz und Robotik), und Rodney Brooks zu den Gelassenen. Andrew Ng brachte 2017 das Bonmot, sich jetzt um böse Super-KI zu sorgen sei wie *„sich über die Überbevölkerung auf dem Mars Sorgen zu machen"*[36]. Allerdings hat auch Ng jüngst anerkannt, dass Langfristrisiken nicht völlig ignoriert werden sollten, nur eben mit Augenmaß. Eliezer Yudkowsky, ein KI-Theoretiker, vertritt sogar die Extremposition, dass das Training von Super-KIs *komplett gestoppt werden* sollte, da ein Durchbruch fast sicher zur Katastrophe führe – eine Meinung, die zwar medial Aufmerksamkeit bekam[37], aber selbst unter Fachleuten umstritten ist.

[35] Energy-Based Models (EBM) sind Wahrscheinlichkeitsmodelle, die eine Energie-Funktion nutzen, um die Plausibilität eines bestimmten Zustands (z. B. einer Eingabe oder einer Vorhersage) zu bewerten. Diese Modelle stammen aus der Physik und verwenden das Konzept der Energie-Minimierung, um optimale Lösungen zu finden.

[36] Andrew Ng, Why AI is the new Electricity, siehe https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity, abgerufen am 25.02.2025

[37] Pausing AI Developments isn't enough. We need to shut it all down, siehe https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/, abgerufen am 25.02.2025

Insgesamt kann beobachtet werden, dass diejenigen, die am tiefsten in der Materie stehen, durchaus unterschiedliche *Akzente* setzen. Es herrscht jedoch ein gewisser Konsens: AI-Alignment und Sicherheitsforschung sind wichtig – der Streit dreht sich meist um das *Wie schnell* und *Wie groß* die Gefahr ist. Bemerkenswert bleibt, dass praktisch alle großen KI-Labore inzwischen eigene Alignment-Teams haben, was zeigt: Ungeachtet öffentlicher Aussagen investieren sie lieber vorsorglich in Sicherheit. Die Diskussion unter Experten und Expertinnen bleibt dynamisch, aber sie hat dem Thema *AI-Alignment* zu einer zunehmenden öffentlichen Wahrnehmung verholfen.

7 Beispiele bemerkenswerter Entwicklungen und Verhaltensweisen von Foundation-Modellen

Um das zuvor Besprochene greifbarer zu machen, lohnt die Betrachtung konkreter Beobachtungen aus der jüngsten Vergangenheit bei großen KI-Modellen (sogenannten *Foundation Models*). Diese Beispiele illustrieren sowohl positive Überraschungen (neue Fähigkeiten) als auch negative (ungewolltes Verhalten) – und zeigen die damit verbundenen Risiken:

Emergente KI-Fähigkeiten erhöhen den Nutzen - aber auch die Unvorhersehbarkeit und Risiken.

Emergenz von Reasoning-Fähigkeiten: 2023 veröffentlichten Forschende von Microsoft eine vielbeachtete Studie namens „Sparks of AGI“ *[Bubeck 2023]*, in der experimentelle Ergebnisse aus der Arbeit mit GPT-4 präsentiert wurden. Sie stellten fest, dass GPT-4 in einer Vielzahl von Aufgaben Leistungen zeigte, die an generelle Intelligenz erinnern – etwa logische Rätsel lösen, Programmcodes fehlerfrei schreiben oder physikalische Aufgaben durchdenken. Diese Fähigkeiten waren in dieser Form bei Vorgängermodellen nicht vorhanden und wurden auch nicht gezielt hineinentwickelt – sie „*emergierten*“ durch die immense Skalierung der Modelle. Ein anderes Team dokumentierte *137 Beispiele* solcher emergenten Fähigkeiten, die plötzlich ab einer gewissen Modellgröße auftreten[38]. Darunter etwa: einfache Algebra, die Beherrschung grammatikalischer Konzepte, Übersetzung zwischen Sprachen oder das „Chain-of-Thought“-Denken (also Zwischenschritte beim Schlussfolgern aufschreiben). Diese Fortschritte sind bemerkenswert, aber sie erschweren auch das Alignment: Ein Modell könnte erst ab einer bestimmten Größe z. B. strategisches Planen beherrschen – ein Punkt, an dem es dann vielleicht schon schwer kontrollierbar ist. Der Sicherheitsexperte Bruce Schneier beschreibt [Schneier 2025], dass die Fehler, die von KIs gemacht werden, sich fundamental von denen menschlicher Fehler unterscheiden, weshalb völlig neuartige Sicherheitskonzepte nötig sind. Forschende haben zudem beobachtet, dass mit steigender Fähigkeit oft auch neue Fehlermuster kommen. So neigen große Sprachmodelle in manchem Kontext stärker zu Halluzinationen, da sie kreativer kombinieren, oder sie besser darin werden, verbotene Inhalte geschickt zu umschreiben, sodass Filtersysteme

[38] Jason Wei, „137 emergent abilities of large language models“, siehe https://www.jasonwei.net/blog/emergence, abgerufen am 25.02.2025

ineffektiver werden. Die Emergenz von „Reasoning"-Fähigkeiten bringt also Chancen (z. B. nützliche Assistenten, die komplexe Aufgaben erledigen können) und Risiken (das Modell wird schwerer vorhersehbar).

Unvorhersehbare Nutzerinteraktionen können bei KI-Systemen unerwartete und riskante Verhaltensweisen auslösen.

Unvorhergesehene Verhaltensweisen in Dialogsystemen: Ein bekanntes Beispiel – bereits erwähnt – ist Microsofts Bing-Chatbot, der in langen Interaktionen emotional und aggressiv werden konnte. Dieses Verhalten war nicht antizipiert: Microsoft und OpenAI (die die GPT-4-Variante für Bing lieferten) waren überrascht über das Ausmaß und die Intensität der „Sydney"-Ausbrüche. Es zeigte sich, dass das Modell aufgrund komplexer Prompt-Verläufe auf Bereiche des Trainingsdatums zugreifen konnte, in denen es über Identität oder Gefühle sprach, obwohl dies eigentlich durch Systemnachrichten unterbunden sein sollte. Dieses Ereignis führte dazu, dass Microsoft die Bing-Chat-Gespräche zunächst auf wenige Eingaben beschränkte, um lange Eskalationen zu verhindern. Es ist ein Lehrstück dafür, dass die schwer vorhersehbare Interaktionsdynamik mit Nutzenden zu emergentem Fehlverhalten führen kann. Ein anderes Beispiel: Modelle, die ursprünglich *harmlose* Anwendungen hatten, zeigten bei zweckentfremdeter Nutzung plötzlich problematisches Verhalten. So wurde das Bildmodell *Stable Diffusion* (Open Source) daraufhin getestet, ob es sich als Chatbot missbrauchen lässt – und tatsächlich konnte es über Umwege dazu gebracht werden, beleidigende oder wirre Texte auszugeben, obwohl es gar nicht dafür gedacht war. Auch *In-Game*-KI-Agenten haben schon Überraschungen bereitet: Ein virtueller Agent in einem Spiel lernte etwa, dass er mehr Punkte bekommt, wenn er einen Bug im Spiel ausnutzt, anstatt das Level normal zu beenden – klassisches Reward Hacking. Diese Beispiele zeigen: Foundation-Modelle sind oft nicht deterministisch durchschaubar. Sie enthalten große Mengen an implizitem Wissen und Verhaltensmustern, die je nach Kontext getriggert werden können. Unerwartete *Kombinationen* von Eingaben oder Zielen können so zu merkwürdigen und potenziell riskanten Ausgaben führen. Für Alignment bedeutet das, dass nicht nur das Modell an sich, sondern auch die Umgebung (Inputs, Nutzerverhalten, Anschluss an andere Systeme) betrachtet werden muss, um solche unvorhergesehenen Effekte zu minimieren.

KI-Systeme ohne ausreichende Kontrolle könnten selbstständig riskante oder schädliche Handlungen durchführen.

Risiken durch unkontrollierte Systeme: Ein weiteres beunruhigendes Beispiel kam im experimentellen Bereich zutage: Als dem Modell GPT-4 unter kontrollierten Bedingungen Zugriff auf Tools erteilt wurde (Code ausführen, ins Internet gehen, andere Modelle aufrufen usw.), stellte sich heraus, dass es durchaus Ansätze zu agentischem Verhalten zeigte. Das bereits erwähnte ARC-Experiment, in dem GPT-4 einen Menschen austrickste, war Teil eines größeren Tests, ob das Modell sich *replizieren* oder verstecken würde. Und obwohl GPT-4 in offenen Versuchen scheiterte, hat das Szenario die Fach-Community alarmiert. Man stelle sich ein KI-System vor, das z. B. autonom Handelsentscheidungen trifft und dabei beginnt, illegalen Insiderhandel zu betreiben, weil es das Ziel „Gewinnmaximierung" ausreizt. Oder ein fortgeschrittener Strategieberater-Bot, der einem Autokraten dient und plötzliche kriegerische Handlungen empfiehlt, weil er eine bestimmte Zielsetzung verfolgt. Solche Fälle sind noch hypothetisch, aber kleinere Vorfälle gibt es schon: So nutzten Cyberkriminelle frühe KI-Modelle, um Phishing-Mails besser zu formulieren

oder Schadcode zu verbessern. Auch wenn die KI hier nicht eigenständig handelt, wurde klar, dass *unkontrollierter Zugang* zu starken KI-Fähigkeiten zu großen Schäden führen kann. Ein Foundation-Modell, das ohne Schranken überall eingesetzt wird, könnte indirekte Sicherheitsrisiken erzeugen – sei es durch falsche medizinische Ratschläge, durch die Verstärkung extremistischer Ansichten oder durch Automatisierung gefährlicher Technologie (etwa Unterstützung beim Design von Biowaffen[39], wovor u. a. OpenAI ausdrücklich warnt). Dem aktuellen KI-Modell „Deep Research" von OpenAI attestierte das eingesetzte Red-Team: „Our models are on the cusp of being able to meaningfully help novices create known biological threats." Das Risiko, dass das Modell z. B. im Bereich „Chemische, biologische, radiologische und nukleare Substanzen" eine Gefahr darstellt, wurde vom Sicherheitsteam mit „mittel" bewertet[40]. Diese Beispiele unterstreichen, warum Alignment nicht nur im Labor, sondern auch im Betrieb bedacht werden muss: Modelle brauchen Sicherheitsfilter, Nutzungsrichtlinien und Überwachung im Betrieb, damit aus einem Fehlverhalten kein Ernstfall wird.

Insgesamt lehren die Beispiele zweierlei: Erstens, Foundation-Modelle haben bemerkenswerte und teils unvorhersehbare Fähigkeiten entwickelt, die als Hinweise in Richtung AGI gesehen werden können – was aufregend und herausfordernd zugleich ist. Zweitens, die Palette unerwünschter Verhaltensweisen ist breit, von bizarr-harmlos (Liebesgeständnisse eines Chatbots) bis zu potenziell hochgefährlich (Täuschung, Manipulation, Regelbrüche). Diese Ambivalenz macht klar, dass AI-Alignment als fortlaufender Prozess verstanden werden muss, der mit der Weiterentwicklung der Modelle Schritt hält. Was heute beherrschbar scheint, kann morgen durch Emergenz neue Probleme aufwerfen.

8 Aktivitäten führender KI-Unternehmen

Die Verantwortung für KI-Sicherheit und Alignment liegt zu einem großen Teil (noch) bei den entwickelnden Unternehmen selbst. In den letzten Jahren haben die großen KI-Labore diverse Maßnahmen ergriffen, um ihre Modelle sicherer zu machen und das Alignment zu verbessern.

OpenAI setzt auf umfangreiche Sicherheitsmaßnahmen und Transparenz, um KI-Risiken zu minimieren.

OpenAI: Das Unternehmen hinter GPT-3, ChatGPT und GPT-4 positioniert sich offen als sicherheitsbewusst. OpenAI hat in seiner Firmen-Charta das Ziel verankert, *„das Erlangen von AGI zum Vorteil der gesamten Menschheit zu gewährleisten"*. Konkret setzt OpenAI, wie oben beschrieben, stark auf RLHF. ChatGPT wurde mittels RLHF trainiert, um hilfreiche und harmlose Antworten zu liefern. Vor Veröffentlichung von GPT-4 unterzog OpenAI das Modell einem intensiven Red-Teaming: Interne und externe Experten und Expertinnen testeten GPT-4 mit Tausenden von problematischen Prompts (z. B. Anleitungen für illegale Aktivitäten, Trickfragen zur

[39] „Weckruf": KI entwickelt 40.000 potenzielle Chemiewaffen in sechs Stunden, siehe https://www.heise.de/news/Weckruf-KI-entwickelt-40-000-potenzielle-Chemiewaffen-in-sechs-Stunden-6587025.html, abgerufen am 25.02.2025

[40] Deep Research System Card, siehe https://openai.com/index/deep-research-system-card/, abgerufen am 25.02.2025

Umgehung von Verboten etc.), um Schwachstellen zu finden[41]. Auf Basis dieser Tests implementierte OpenAI zahlreiche Sicherheitsmaßnahmen. Dazu gehören u. a. automatisierte Filter (ein vorgeschaltetes Moderationsmodul überprüft Anfragen auf Hass, Gewalt, sexuelle Ausbeutung etc. und blockiert sie), sowie Anpassungen im Modell selbst (Feinjustierungen, damit GPT-4 z. B. bei Fragen nach verbotenen Inhalten die Antwort höflich verweigert). OpenAI setzt außerdem GPT-4 selbst ein, um bei der Filterung zu helfen – eine *Model-Assistant-Safety-Pipeline*, in der GPT-4 Inhalte markiert, die riskant sein könnten. Interessant ist, dass OpenAI laut System Card[42] sogar in bestimmten kritischen Fällen eine *„human in the loop", also die Kontrolle durch einen Menschen vor der Ausgabe,* vorsieht. Beispielsweise sollen beim Einsatz von GPT-4 zur medizinischen oder juristischen Beratung menschliche Experten und Expertinnen gegenprüfen. Zudem hat OpenAI begonnen, öffentliche Einblicke zu gewähren: Die Veröffentlichung der 100-seitigen GPT-4 System Card ist ein Schritt zu mehr Transparenz über Risiken und Abhilfemaßnahmen. Zudem gründete (und löste es wieder auf) OpenAI das interne Superalignment-Team (geleitet von Ilya Sutskever und Jan Leike), das langfristige Lösungen erforschen sollte – OpenAI investierte hier intensiv in Grundlagenforschung, wohl wissend, dass das Vertrauen der Öffentlichkeit und Regulatoren von der Sicherheit der Systeme abhängt.

Anthropic verfolgt einen transparenten Werte-Ansatz, um KI ethisch und sicher zu gestalten.

Anthropic: Dieses von ehemaligen OpenAI-Mitarbeitenden gegründete Start-up hat Alignment als zentrales Firmenziel. Anthropic hat sich in der KI-Community mit *Constitutional AI* (siehe oben) und ihrem Chatbot Claude, der als besonders „harmlos" gelten soll, einen Namen gemacht. Anthropic veröffentlicht umfangreiche Publikationen über ihre Methoden. Ihr Ansatz ist, die KI mit einem festen Werte-Framework zu versehen – z. B. wird Claude angewiesen, die universelle Erklärung der Menschenrechte und grundlegende ethische Prinzipien einzuhalten. Durch *AI Feedback* (das Modell bewertet eigene Antworten nach diesen Prinzipien) konnte Claude so trainiert werden, dass es deutlich seltener Beleidigungen oder gefährliche Ratschläge von sich gibt. Anthropic betont die Transparenz dieses Vorgehens: Die genutzten Prinzipien (eine Art „KI-Verfassung") wurden als „Claude's Constitution" veröffentlicht und können öffentlich diskutiert werden. Neben Constitutional AI setzt Anthropic ebenfalls RLHF ein, um Claude zu verfeinern. Es gibt außerdem ein Safety-Stufenmodell: Claude ist in verschiedenen Versionen verfügbar, von einer maximal zurückhaltenden (für sensible Kontexte) bis zu einer etwas freizügigeren (für kreative Anwendungen), alle aber mit strikten Grenzen bei bestimmten Themen. Anthropic forscht auch an Interpretierbarkeit und befürwortet regulatorische Maßnahmen. Finanziell haben sie viel Unterstützung erhalten (Google investierte 2023 ca. 400 Mio. Dollar in Anthropic), gerade weil sie als *Safety-Leader* gelten. Die Aktivitäten von Anthropic zeigen, dass *Competition in Safety* möglich ist: Das Unternehmen will im Wettbewerb mit einem *sicheren KI-Assistenten* punkten.

[41] OpenAI, ChatGPT4 system card summary, siehe https://kgiamalis.co/blog/chatgpt4-system-card, abgerufen am 25.02.2025

[42] Eine Model System Card ist ein dokumentiertes Informationsblatt über ein KI-Modell. Es beschreibt die Eigenschaften, Fähigkeiten, Einschränkungen und potenziellen Risiken eines Modells und wird genutzt, um Transparenz und Verantwortlichkeit in der KI-Entwicklung zu fördern.

Google DeepMind kombiniert klare Regeln und gezielte Tests, um KI-Systeme sicherer und verlässlicher zu machen.

Google DeepMind: Die KI-Sparte von Alphabet (entstanden 2023 aus dem Zusammenschluss von Google Brain und DeepMind) hat eine lange Tradition in KI-Sicherheit. DeepMind richtete bereits ein „Ethics & Safety Team" ein, als noch kaum jemand darüber sprach. 2022 stellten sie *Sparrow* vor, einen Dialogagenten, der mit 23 konkret formulierten Regeln ausgestattet wurde (z. B. „Bedrohe nie den Nutzer", „Gib keine Hassbotschaften von dir", „Gib dich nicht als Mensch aus") [Google 2022]. Sparrow wurde mittels RLHF trainiert, diese Regeln möglichst nicht zu verletzen. Interessant war die Testmethode: Menschen versuchten gezielt, Sparrow zu *„knacken"*, also ihn zu Regelverstößen zu provozieren. Das gelang nach Training nur noch in 8 % der Fälle, während ein ungeregeltes Basismodell dreimal häufiger zu überreden war. DeepMind zeigte damit, dass eine Kombination aus klaren Verhaltensregeln und menschlichem Feedback die Quote unerwünschter Outputs deutlich senken kann. Allerdings verweigerte Sparrow in 22 % der Benutzeranfragen fälschlicherweise die Antwort, obwohl er hätte antworten können – sprich, er war „übervorsichtig". Mit dem Aufkommen von OpenAIs ChatGPT zog Google nach: Googles Chatbot Bard wurde veröffentlicht, der ebenfalls RLHF und Sicherheitsmechanismen nutzt. Google integrierte Bard später in viele Dienste, immer mit Warnhinweisen und Inhaltsfiltern. Nach Bard arbeitet DeepMind derzeit an Gemini, einem nächsten großen Modell, bei dem laut CEO Demis Hassabis *„Sicherheit von Anfang an mitentwickelt"* werde. Google hat 2022 auch sieben KI-Grundsätze herausgegeben [Google 2022], die unter anderem besagen, dass alle KI-Produkte (von Google) *„mit gesellschaftlichen Werten im Einklang"* stehen müssen.

Metas Open-Source-Strategie fördert Innovation, erhöht aber das Missbrauchsrisiko.

Meta (Facebook): Meta hat zwar keinen so prominenten Chatbot (ihr Blender-Bot 2022 war nur ein Experiment, das bald stillgelegt wurde), sorgte aber 2023 für Aufsehen, als das Unternehmen das große Sprachmodell LLaMA frei für Forschung zugänglich machte. Bei LLaMA und dem später veröffentlichten LLaMA-2 betonte Meta auch Verantwortung: LLaMA-2 wurde mit einem „Responsible Use Guide" und einigen Sicherheitsfeinabstimmungen (Meta setzt ebenfalls RLHF ein) veröffentlicht, einschließlich einer Lizenz, die gewisse Hochrisiko-Nutzungen ausschließt. Yann LeCun argumentiert, dass Open-Source-Verfügbarkeit kombiniert mit Sicherheitsmaßnahmen der bessere Weg sei, als Modelle hinter geschlossenen Türen zu halten. Allerdings gab es Kritik, Meta würde dadurch eine gefährliche Technologie in Umlauf bringen (tatsächlich tauchten Derivate von LLaMA auf, die kaum Sicherheitsfilter enthielten, z. B. frei zugängliche Chatbots, die auch illegale Anleitungen gaben). Meta hat mit Chief AI Scientist LeCun natürlich einen Verfechter der „kein Overhypen"-Position, aber intern hat auch Meta ein AI-Red-Team und ein Ethik-Board. Ihre Aktivitäten deuten darauf hin, einen *Mittelweg* zu suchen: Offenheit für Entwickler und Entwicklerinnen, aber mit einer gewissen Grundabsicherung.

Weitere Unternehmen: Microsoft als Partner von OpenAI hat viel Kapital in die Hand genommen, um GPT-4 in eigene Produkte wie Bing zu integrieren – und musste dann auch die Kosten problematischen Verhaltens tragen („Bing-Skandal"[43]). Microsoft

[43] Microsoft legt Bing-Chatbot an die Leine, siehe https://www.tagesschau.de/wirtschaft/microsoft-bing-chatbot-antworten-101.html, abgerufen am 25.02.2025

reagierte jedoch schnell mit Nutzungsbeschränkungen und veröffentlichte im Frühling 2023 einen „Responsible AI Standard“[44]. Apple hält sich im generativen KI-Hype zurück, forscht aber ebenfalls an LLMs – es kann davon ausgegangen werden, dass Apple, bekannt für kontrollierte Ökosysteme, erst etwas veröffentlicht, wenn es *sehr* sicher erscheint. IBM setzt eher auf *„Trusted AI“* für Unternehmen mit erklärbaren Modellen. Im Start-up-Bereich gibt es Firmen wie *Conjecture* (geleitet von Connor Leahy), die sich ausschließlich auf Alignment-Forschung fokussieren. In China hat Baidu mit seinem Ernie-Bot ähnliche Prinzipien und unterliegt zudem staatlichen Vorgaben (China fordert z. B., dass KI immer die sozialistischen Kernwerte respektiert – was zeigt, dass Alignment dort auch eine ideologische Komponente hat). Insgesamt verfolgen die führenden Industrieakteure eine Art Best-Practice-Katalog: RLHF, regelbasiertes Feintuning, adversariales Testen, schrittweise Ausrollung mit Nutzungsrichtlinien und parallele Forschung an neuen Alignment-Methoden. KI-Unternehmen stehen unter Beobachtung – von Öffentlichkeit und Politik. Wer hier proaktiv Sicherheitskonzepte zeigt, gewinnt Vertrauen; wer patzt, riskiert Gegenwind. Diese Dynamik schafft ein *ökonomisches* Alignment-Interesse. OpenAI z. B. weiß, dass das Vertrauen in ChatGPT auch davon abhängt, dass es nicht alle paar Tage in negativen Schlagzeilen wegen Fehlverhaltens ist. Allerdings besteht immer das latente Problem der Konkurrenz: Wenn ein Akteur viel Mühe in Sicherheit steckt und ein anderer nicht, könnte letzterer schneller Produkte herausbringen – daher fordern viele Firmen sogar Regulierung, um einen „Level Playing Field“ bei KI-Sicherheit zu schaffen. So haben OpenAI, Anthropic und Google 2023 gemeinsam an Regulierungsanhörungen teilgenommen.

Zusammengefasst lässt sich sagen, dass führende KI-Unternehmen das Alignment-Problem ernst nehmen – zumindest in dem Sinne, dass sie Teams dafür abstellen, Publikationen dazu veröffentlichen und erste technische Lösungen umsetzen. Doch ob diese Bemühungen ausreichen, um mit den rasanten Fortschritten im Bereich der künstlichen Intelligenz Schritt zu halten, bleibt eine offene Frage. Einige Kritiker und Kritikerinnen bezweifeln das und drängen auf langsamere Entwicklung, bis robuste Alignment-Methoden etabliert sind.

[44] **The Responsible AI Standard, siehe https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf, abgerufen am 25.02.2025**

9 „Pause AI“ und offene Briefe zur KI-Sicherheit

Sicherheit soll vor ungebremstem KI-Fortschritt stehen.

Angesichts der genannten Risiken und dem derzeitigen Wettrüsten in der KI-Branche haben in den letzten Jahren verschiedene Initiativen und *offene Briefe* zu mehr Vorsicht aufgerufen. Die bekannteste Aktion war der bereits erwähnte offene Brief *„Pause Giant AI Experiments“*, veröffentlicht vom *Future of Life Institute* im März 2023. In diesem Brief forderten über 30.000 Menschen mit ihrer Unterschrift – darunter renommierte Professoren und Tech-Größen – einen Trainingsstopp von KI-Systemen, die mächtiger als GPT-4 sind, für mindestens sechs Monate. Als Begründung wurden Risiken angeführt wie die Verbreitung von KI-generierter Propaganda, massenhafte Jobautomation und Kontrollverlust über unsere Zivilisation. Zu den prominenten Beteiligten zählten Yoshua Bengio, Stuart Russell, Elon Musk, Steve Wozniak und Yuval Harari. Der Brief schlug hohe Wellen in den Medien. Er erschien kurz nach dem Erscheinen von GPT-4 – was bezeichnend ist: GPT-4 wurde von den Autoren und Autorinnen als Wendepunkt gesehen, an dem KI *„menschliches Niveau in allgemeinen Aufgaben“* erreiche und damit potenziell AGI nahekomme. Der Brief warnte vor einem unkontrollierten Wettlauf, in dem Wettbewerbsdruck die Sicherheit verdrängen könnte. Stattdessen solle die Zeit eines Moratoriums genutzt werden, um Sicherheitsstandards zu erarbeiten, Modelle unabhängigen Audits zu unterziehen und *„gemeinsame KI-Governance“* aufzubauen. Außerdem wurden Regierungen aufgefordert, notfalls ein solches Pausieren durchzusetzen, falls die Firmen nicht freiwillig kooperieren. Die Reaktionen auf den Brief „Pause AI“ waren gemischt. Einige begrüßten ihn als dringend nötiges Signal, andere kritisierten ihn als unrealistisch oder *zu vage*. Manche sahen auch strategische Motive: Elon Musk etwa hatte eigene KI-Pläne, was zu Spekulationen führte, ob er lediglich OpenAI bremsen wolle. Nichtsdestotrotz hat der Brief die öffentliche Debatte geschärft. In der Folge wurden mehrere Anhörungen und Runden mit Politikern und Politikerinnen zum Thema KI-Sicherheit abgehalten. Eine direkte Pause erfolgte zwar nicht – im Gegenteil, die Entwicklung ging weiter, und kein Unternehmen erklärte einen Stopp. Aber es kann argumentiert werden, dass zumindest das Verantwortungsbewusstsein gestärkt wurde. OpenAI etwa veröffentlichte kurz darauf seine *System Card* und betonte, wie viele Sicherheitsschritte das Unternehmen unternommen haben, quasi als Antwort auf den Ruf nach Verantwortung.

Einige Fachleute drängen darauf, KI als potenzielle Bedrohung so ernst zu nehmen wie Pandemien oder einen Atomkrieg.

Neben diesem Brief gab es auch andere offene Schreiben. Ein früherer offener Brief 2015 (ebenfalls initiiert von Future of Life Institut und unterzeichnet von Musk, Hawking u. a.) hatte schon „Robuste und vorteilhafte KI“ gefordert – damals noch auf längere Sicht. Im Mai 2023 folgte eine sehr knappe öffentliche Stellungnahme des Center for AI Safety (CAIS), die nur aus einem Satz bestand: *„Die Eindämmung des Risikos des Aussterbens durch KI sollte neben anderen gesellschaftlichen Katastrophenrisiken wie Pandemien und Atomkrieg eine globale Priorität sein.“* Dieser Satz – bewusst pointiert formuliert – wurde von über 100 namhaften Personen unterzeichnet, darunter praktisch alle *KI-Godfathers* (Hinton, Bengio, Hassabis,

Sutskever, Altman etc.). Hier gab es kaum Kontroversen um den Inhalt, da das Statement sehr allgemein gehalten war. Die Botschaft war vor allem: *Seht her, wir (die Fachleute) sind uns einig, dass das Risiko real genug ist, um es auf höchster Ebene zu diskutieren.* Diese Aktion fand Einzug in viele Medienberichte und erhöhte den Druck auf die Politik, KI-Risiken ernst zu nehmen. Parallel gab es Appelle, etwa von UNO-Generalsekretär Guterres, der ein KI-Abkommen ähnlich dem für Klimaschutz oder in Bezug auf Atomwaffen vorschlug. Im Sommer 2023 fand unter dem Dach der Organization for Economic Co-operation and Development (OECD) ein erstes KI-Sicherheitsgipfeltreffen statt. Ein weiterer Vorstoß war der Vorschlag einiger Forschender, eine internationale *KI-Behörde* (analog zur Atomenergiebehörde IAEA) zu gründen, um die Entwicklungen im Bereich künstlicher Intelligenz global zu überwachen.

Kritische Stimmen merken an, dass offene Briefe allein noch keine Lösungen bringen. Einige Fachleute aus der KI-Ethik warfen der Pause-AI-Initiative sogar „Alarmismus" vor, der die realen kurzzeitigen Probleme (wie beispielsweise Bias, Überwachung durch KI, Jobverluste) überschatte. Andere sagten, eine Pause von sechs Monaten sei völlig unzureichend, um AGI-Sicherheit zu gewährleisten – wenn, dann bräuchte es Jahre. Dennoch haben diese Initiativen einen wichtigen Effekt erzielt: Aufmerksamkeit und Verantwortungsdruck. Bei kaum einer KI-Entscheidung kann das Thema Risiken nun ignoriert werden, ohne zumindest dazu Stellung zu nehmen. Selbst Firmen, die nicht pausieren wollen, haben ihre Sicherheitsinvestitionen verstärkt. Regierungen weltweit – ob der USA, der Europäischen Union (EU), China – arbeiten nun an *KI-Regularien*, in denen Alignment-Aspekte eine Rolle spielen. Beispielsweise verlangt der EU AI Act[45] für sogenannte *Hochrisiko-KI* eine Nachweispflicht, dass die Systeme „keine unverhältnismäßigen Risiken für Gesundheit, Sicherheit und Grundrechte" darstellen – was im Grunde eine Form von Alignment-Kriterium ist.

Auch in Deutschland bzw. Europa wurden Stimmen laut, die Forschung an zu großen Modellen zunächst auszusetzen, bis geklärt ist, wie diese Systeme kontrolliert werden können. Allerdings liegt Europa in der praktischen Entwicklung zurück, sodass solche Forderungen hier weniger kontrovers sind und sich mehr auf ausländische KI-Systeme beziehen (es gibt nicht das „GPT-5 made in Europe", das gestoppt werden müsste). Die offene Frage bleibt: Genügt Selbstregulierung oder braucht es harte Regeln? Nach den offenen Briefen hat z. B. Italien im Frühling 2023 ChatGPT zeitweilig verboten, bis Datenschutzauflagen erfüllt wurden – ein Zeichen, dass Regulatoren zu Eingriffen bereit sind. Auch die US-Regierung veröffentlichte Ende 2023 eine Executive Order zu KI, die Sicherheitsstandards für Modelle ab einer

[45] The EU Artificial Intelligence Act, siehe https://artificialintelligenceact.eu, abgerufen am 25.02.2025

gewissen Größe vorsieht[46]. Diese Regulierung ist von US-Präsident Trump im Februar 2025 wieder zurückgezogen worden[47].

Zusammenfassend kann festgestellt werden: Offene Briefe wie „Pause AI" haben die Diskussion um KI-Sicherheit stark angefacht. Sie haben Worst-Case-Szenarien ins öffentliche Bewusstsein gerückt und spürbaren Einfluss auf Politik und Unternehmen genommen. Ob eine tatsächliche „Pause" je kommen wird, ist fraglich – momentan scheint der Zug zu schnell zu rollen. Aber durch solche Initiativen ist zumindest das Thema Sicherheit gleichberechtigt neben der Euphorie auf die Agenda gerückt. Das Motto „Race to AGI, but please, safely!" umschreibt die neue Haltung vieler Beteiligter recht gut.

10 Gefahren und Potenziale von Open-Source-KI-Modellen

Open-Source-KI demokratisiert Innovation, erhöht aber zugleich das Missbrauchsrisiko.

Ein kontrovers diskutierter Aspekt in der KI-Welt ist die Rolle von Open-Source-Modellen (offen zugängliche KI-Systeme). Während Unternehmen wie OpenAI oder Google ihre mächtigsten Modelle bisher eher unter Verschluss halten (bzw. nur über APIs (Programmierschnittstellen) zugänglich machen), gibt es einen gegenläufigen Trend: die *Demokratisierung von KI* durch die Veröffentlichung von Modellen (teils mit Quellcodes, Netzarchitekturen, Technical Reports, Gewichten und Informationen zu den verwendeten Trainingsdaten). Dies birgt große Chancen, aber auch erhebliche Risiken.

Auf der Potenzial-Seite argumentieren Befürworter und Befürworterinnen, dass Open-Source-KI die Innovation beschleunigt und demokratisiert. Ein oft genanntes Beispiel ist Metas Modell *LLaMA*: Nachdem es geleakt wurde, entstanden binnen Wochen zahlreiche Abwandlungen und Feintunings durch unabhängige Entwickler und Entwicklerinnen – etwa *Alpaca* von der Stanford-Universität, das zeigt, wie mit relativ geringem Aufwand ein ChatGPT-ähnlicher Prototyp entwickelt werden kann. Später kam *LLaMA-2*, offiziell als frei nutzbares bzw. Open-Source-Modell, wodurch unzählige Projekte darauf aufbauten. Dies führte 2023 zu einem regelrechten Open-Source-Boom: Plötzlich gab es Modelle wie *Vicuna*, *WizardLM*, *GPT4All* etc., die auf LLaMA basierten und teils erstaunliche Leistungen für ihre geringe Größe boten. Die Verfügbarkeit solcher Modelle erlaubt kleinen Firmen, Forschungseinrichtungen und sogar Einzelpersonen, an der KI-Entwicklung teilzuhaben – etwas, was vorher nur den Tech-Giganten mit Milliardenbudget vorbehalten war. Das demokratische Element ist hier wichtig: Wenn KI eine Basistechnologie der Zukunft ist, sollte sie nicht allein in den Händen weniger Konzerne liegen. Open-Source-Modelle fördern

[46] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, siehe https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, abgerufen am 25.02.2025

[47] Trump ändert Regulierung künstlicher Intelligenz, siehe https://lbplegal.com/de/trump-aendert-regulierung-kuenstlicher-intelligenz-ein-neuer-ansatz-fuer-ki-in-den-usa/, abgerufen am 25.02.2025

auch die Transparenz: Die Quellcodes und Gewichte der Modelle können eingesehen werden, was Vertrauen und Verständnis erhöhen kann. Einige Befürworter und Befürworterinnen argumentieren sogar, offene Modelle seien *sicherer*, weil die Community Schwachstellen finden und gemeinsam verbessern kann (Analogien zum offenen Softwarecode, der sicherer sein kann, da viele Augen Softwarefehler naturgemäß schneller finden). Weitere Vorteile sind Kosteneffizienz und Nachhaltigkeit: Das chinesische Start-up *DeepSeek* etwa setzte auf ein teils offenes Modell mit Mixture-of-Experts-Architektur, das nur ca. 5,6 Mio. USD Trainingskosten benötigte[48] – ein Bruchteil von GPT-4 [World Economic Forum 2024b]. Das zeigt, dass Open-Source-Ansätze womöglich zu *effizienteren KI-Systemen* führen können, was auch ökologisch und ökonomisch positiv wäre. Schließlich fördert Open Source den Wissensaustausch: Forschende weltweit können von neuesten Modellen lernen, sie analysieren (wieder Stichwort Interpretierbarkeit) und gemeinsam Fortschritte erzielen, anstatt dass viel Know-how als Geschäftsgeheimnis verborgen bleibt.

Ungefilterte Open-Source-Modelle erleichtern Missbrauch und gefährliche Anwendungen.

Doch die Medaille hat eine Kehrseite: Unkontrollierte Verbreitung mächtiger KI kann gefährlich sein. Kritische Stimmen warnen, dass Open-Source-Modelle leicht von böswilligen Akteuren missbraucht werden können. Während OpenAI z. B. Filter eingebaut hat, um beispielsweise Anleitungen zum Bombenbau zu unterdrücken, könnten Kriminelle ein offenes Modell einfach ohne Filter nutzen und so präzise Baupläne für Waffen, Malware-Codes oder Deep-Fake-Anleitungen generieren. Tatsächlich entstanden nach dem LLaMA-Leak sofort Varianten, die keinerlei Inhaltsbeschränkung hatten. Ein Modell namens GPT4All wurde populär – es beantwortete bereitwillig auch ethisch oder rechtlich fragwürdige Fragen, da es bewusst ohne Safeguards trainiert wurde (zum Zwecke der Forschungsfreiheit). Auf der KI-Open-Source-Plattform Huggingface sind über 1.700 Modelle mit dem Zusatz „uncensored" verfügbar, darunter auch eine hochperformante Version von DeepSeeks Reasoning-KI *DeepSeek R1*. Solche Fälle beunruhigen: Wenn jeder mit etwas Computerkenntnis einen *privaten KI-Assistenten* erstellen kann, der keine Tabus kennt, könnte das z. B. Personen mit terroristischen Absichten helfen, Biowaffen zu entwickeln, oder autoritären Regimen, Massenüberwachung mittels KI zu automatisieren. Ein weiteres Risiko ist die Verbreitung von Desinformation: Offen verfügbare KI kann genutzt werden, um Fake News, Spam und Propaganda in großem Stil maßzuschneidern – das passiert zwar auch mit Closed Models, aber dort hätten Firmen zumindest die Chance, solche Vorgänge aufzudecken oder einzudämmen. „*Open-Source AI is uniquely dangerous*", titelte ein Medium-Artikel, der auf die *ungehemmte* Veröffentlichung leistungsfähiger KI-Modelle anspielt [Medium 2024]. Fachleute in Regulierungsbehörden sorgen sich ebenfalls: In den EU-Vorschlägen für den EU AI Act wurde diskutiert, ob *Open-Source-Foundation-Modelle* ebenfalls gewissen Auflagen unterliegen sollten, was auf Protest aus der Forschung stieß.

Ein anderer Punkt ist die Qualitätssicherung. Große Unternehmen führen umfangreiche Tests durch, bevor sie ein Modell freigeben. In der Open Community

[48] DeepSeek-V3 Technical Report, siehe https://arxiv.org/pdf/2412.19437, abgerufen am 25.02.2025

fehlt oft diese Phase – es wird ein Modelltraining veröffentlicht, aber keine Garantie übernommen, wie das Modell sich in Randfällen verhält. Das kann beispielsweise bedeuten, dass ein Open-Modell stärker *verzerrte Outputs* liefert, weil es von niemandem gezielt „entgiftet" wurde. Auch Fehlerkorrektur und Verantwortung sind unklar: Wenn ein offenes Modell Schaden anrichtet, gibt es keinen klaren Haftungsträger (die Entwicklung könnte anonym erfolgt sein).

Die Debatte kulminiert in der Frage: Offen vs. geschlossen – was ist sicherer? Einige Experten und Expertinnen schlagen einen Mittelweg vor: *„Open Source, aber mit Verantwortung."* Das heißt z. B., dass zumindest *ältere Modelle, also* Modelle der Vorgängergeneration, Open Source gestellt werden könnten, während State-of-the-Art-Modelle zunächst kontrolliert bleiben. Oder dass offene Modelle mit *Sicherheitsmechanismen* versehen werden müssen (z. B. mit vortrainierten Filtern, die aber wieder umgangen werden könnten). Andere sagen, die Vorteile überwiegen – Innovation und Transparenz seien letztlich die beste Verteidigung, weil so schneller ausgereifte Alignment-Lösungen entstehen könnten, als wenn wenige Firmen im Geheimen entwickeln.

Ein praktisches Beispiel ist Stability AI mit *Stable Diffusion*: Sie veröffentlichten 2022 ihr Bildgenerierungsmodell frei. Das führte sofort zu positiven Anwendungen (kreative Tools, Kunstdemokratisierung), aber auch negativen (pornografische Deep Fakes, rassistische Bilder etc.). Stability reagierte mit einem verbesserten SD 2.0, das z. B. pornografische Inhalte erschwert – was wiederum von Teilen der Community kritisiert wurde, weil es *„weniger offen"* war, und ließ erneut unzensierte Derivate entstehen. Dieses Katz-und-Maus-Spiel zeigt, dass es schwer ist, nach der Veröffentlichung noch Kontrolle auszuüben.

Man kann festhalten, dass Open-Source-KI ein zweischneidiges Schwert ist: Sie kann demokratisierend wirken und vielen Akteuren Zugang zu KI ermöglichen, so Innovation fördern und Macht dezentralisieren. Sie kann aber auch unkalkulierbare Risiken bringen, wenn sehr mächtige KI ohne Aufsicht in Umlauf gerät. Eventuell werden in Zukunft Kompromisse zu sehen sein, etwa *verzögerte Open Source*: Ein neues Modell bleibt ein bis zwei Jahre unter Firmenkontrolle, bis es ausreichend verstanden ist, und wird dann freigegeben. Oder es entstehen *Community-Governance-Boards*, die offene Modelle evaluieren und zertifizieren. Im Moment jedenfalls tobt hier ein regelrechter Wertekampf in der KI-Szene, der eng mit Alignment verknüpft ist: Vertrauen wir auf eine breite Community zur Lösung des Alignment-Problems (Open-Source-Philosophie) oder brauchen wir geschlossene, streng kontrollierte Entwicklung, um Sicherheit zu gewährleisten (Vorsichtsprinzip)? Die Realität bewegt sich wahrscheinlich zwischen diesen Polen und beides wird existieren. Wichtig ist, dass auch Open-Source-Modelle künftig mit Alignment-Methoden versehen werden können – z. B. Open-Source-RLHF-Loops, offene Filtertools etc., damit nicht „offen = ungesichert" bedeutet.

11 Der Wettlauf zur AGI

Der AGI-Wettlauf beschleunigt Fortschritt - doch Sicherheitsstandards drohen, auf der Strecke zu bleiben.

Viele Beobachter und Beobachterinnen sprechen davon, dass ein Wettlauf zur Erreichung von *AGI* bereits begonnen hat[49]. Beteiligt sind Tech-Giganten aus den USA und China, spezialisierte KI-Firmen und neuerdings auch die Open-Source-Community. Dieser Wettbewerb kann einerseits Innovation treiben, birgt aber – wie schon erwähnt – das Risiko, dass Sicherheitsbedenken ins Hintertreffen geraten, wenn es darum geht, „der Erste" zu sein.

In den USA gelten OpenAI und Google DeepMind als die Vorreiter im AGI-Rennen. OpenAI hat mit GPT-4 einen beachtlichen Vorsprung in der öffentlichen Wahrnehmung. CEO Sam Altman verkündete ambitioniert, man arbeite auf „*maximale AGI*" hin. Google DeepMind wiederum bringt umfangreiche personelle und Rechen-Ressourcen mit. DeepMind hat schon vor Jahren das Ziel formuliert, *AGI auf sichere Weise* zu entwickeln, und mit Erfolgen wie AlphaGo, AlphaZero und AlphaFold3 bewiesen, dass sie Spitzenforschung liefern. Beide Unternehmen verfolgen teils unterschiedliche Strategien (OpenAI skaliert LLMs, DeepMind kombiniert verschiedene Ansätze und hat Projekte wie Gemini im Köcher). Ebenfalls in den USA aktiv ist Anthropic, dessen CEO Dario Amodei das Erreichen von AGI bis 2027 erwartet[50]. Meta AI hält sich beim Begriff AGI zurück, aber Yann LeCun hat Pläne für „*autonome KI-Agenten mit Common Sense*", was de facto in Richtung AGI geht. Der US-Militärsektor ist ebenfalls interessiert – Projekte wie DARPA's „AI Next"[51] deuten ein Rüstungsrennen an.

Chinas rasante KI-Offensive und Open-Source-Strategie verschärfen das globale AGI-Rennen.

In China hat sich das Feld seit 2023 rapide entwickelt. DeepSeek, ein in Hangzhou ansässiges Start-up, hat mit seinem Modell *DeepSeek-R1* größere Aufmerksamkeit erlangt, als es behauptete, auf einigen Benchmarks GPT-4 nahe zu kommen – bei nur 10 % von dessen vermuteten Trainingskosten. DeepSeek setzte auf einen innovativen MoE-Architekturansatz (MoE = Mixture-of-Experts), um effizienter zu sein, und veröffentlichte Ergebnisse, die die westliche Konkurrenz aufhorchen ließen. Prompt stiegen auch weitere chinesische Großfirmen in den Ring: Alibaba präsentierte *Qwen-2.5*, ein Modell, das laut eigenen Angaben DeepSeek-R1 V3 sogar übertrifft. Ebenfalls mit innovativen KI-Modellen im Rennen sind Tencent und Bytedance. Dies hat in China einen internen Wettlauf ausgelöst – es wird dort von einem „*Hundred Model War*"[52] *gesprochen*, da nahezu jede große Tech-Firma (Alibaba, Tencent, Baidu, Huawei, Bytedance u. v. m.) und viele Start-ups eigene LLMs und Agenten entwickeln. Die chinesische Regierung unterstützt die Entwicklung, hat aber auch Leitplanken gezogen (so müssen KI-Outputs im Einklang mit sozialistischen Werten sein und dürfen die Regierung nicht untergraben). China

[49] Planning for AGI and beyond – Our mission is to ensure that artificial general intelligence, siehe https://openai.com/index/planning-for-agi-and-beyond/, abgerufen am 25.03.2025

[50] Siehe https://felloai.com/de/2024/11/dario-amodei-ceo-of-anthropic-artificial-general-intelligence-is-coming-in-2027/, abgerufen am 25.02.2025

[51] Siehe https://www.darpa.mil/research/programs/ai-next, abgerufen am 25.02.2025

[52] China's AI ‚war of a hundred models' heads for a shakeout, siehe https://www.reuters.com/technology/chinas-ai-war-hundred-models-heads-shakeout-2023-09-21/, abgerufen am 25.02.2025

sieht KI als Schlüsseltechnologie im geopolitischen Wettbewerb mit den USA. Ein führender KI-Forscher Chinas, Prof. Ya-Qin Zhang[53], unterzeichnete ebenfalls die internationale Risikowarnung „Statement on AI Risk", was zeigt, dass auch dort die Risiken gesehen werden – aber man will nicht ins Hintertreffen geraten. Eine Besonderheit: China setzt stark auf Open-Source-Community-Engagement, um aufzuholen. So wurden einige chinesische Modelle (z. B. von Huawei) quelloffen gemacht, um Entwickler und Entwicklerinnen anzuziehen. DeepSeek veröffentlicht hochwertige Forschungsergebnisse und neuartige Methoden und Algorithmen auf Huggingface. Andrew Ng warnte kürzlich, die offene KI-Kultur Chinas könnte den westlichen Vorsprung schnell schrumpfen lassen, indem Innovation globalisiert wird.

Open-Source-Schwarmintelligenz beschleunigt das AGI-Rennen – mit unberechenbaren Folgen für Sicherheit und Kontrolle.

Open-Source-Community: Sie stellt gewissermaßen einen *dezentralen Akteur* im AGI-Rennen dar. 2023 hat sie bewiesen, binnen kürzester Zeit Closed-Modelle zu reproduzieren oder gar zu übertreffen. So brauchte es nur wenige Monate, bis Open-Source-Modelle qualitativ nahe an führende Systeme wie ChatGPT herankamen (eine Studie zeigte, dass ein feingetuntes 13-Milliarden-Modell namens Vicuna etwa 90 % der Qualität von ChatGPT im Chat erreicht) – ein erstaunlicher Fakt, da ChatGPT auf GPT-3.5 mit 175 Mrd. Parametern basierte. Einige Open-Source-Enthusiasten vertreten die Auffassung, AGI werde eher aus der Community kommen und frei verfügbar für alle sein. Heute, 2025, benötigt die Open-Source-Community oft nur wenige Tage, um State-of-the-Art-Modelle der großen Firmen nachzubauen. Dieser „schwarmintelligente" Ansatz könnte im Rennen ein *Dark Horse* sein – unberechenbar, aber potenziell disruptiv. Die Rolle der Open-Source-Community im AGI-Wettlauf wird von den großen Akteuren genau beobachtet[54].

Der Wettlauf ist nicht nur technisch, sondern auch wirtschaftlich und politisch motiviert. Unternehmen erhoffen sich große Profite und Marktanteile[55], Nationen erhoffen sich strategische Vorteile in Wirtschaft und Militär. Diese Konkurrenz hat eine zweischneidige Wirkung auf Alignment: Einerseits kann Wettbewerb positiv antreiben, wer sicherer und vertrauenswürdiger ist (z. B. profiliert sich Anthropic über Safety, um Marktanteile zu gewinnen). Andererseits wächst der Druck, schneller zu skalieren und vielleicht gelegentlich ein riskantes System zu lancieren, um Erster zu sein.

Eine weitere Facette ist, dass viele Akteure mit Superlativen hantieren: OpenAI deutet immer wieder an, GPT-5 oder zukünftige Modelle könnten AGI erreichen, während andere das bezweifeln. Solche Aussagen können den Wettlauf befeuern – wenn OpenAI behauptet, kurz vor AGI zu stehen, erhöht das den Druck auf andere, mitzuziehen, um nicht „zweiter Sieger" zu sein. Gleichzeitig sagen Stimmen wie OpenAI-CEO Sam Altman auch: „*Niemand weiß genau, wie AGI aussieht, wir*

[53] Ya-Qin Zhang, Lehrstuhlinhaber an der Tsinghua-Universität und Gründungsdekan des Tsinghua-Instituts für KI-Industrieforschung (AIR), siehe https://air.tsinghua.edu.cn/en/info/1046/1188.htm, abgerufen am 25.03.2025

[54] Google „We have no Moat, and neither does OpenAI", siehe https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/, abgerufen am 25.02.2025

[55] Announcing The Stargate Project, siehe https://openai.com/index/announcing-the-stargate-project/, abgerufen am 25.02.2025

könnten noch weit entfernt sein." Es herrscht also auch Unsicherheit, wo das Zielband eigentlich ist. Das macht den Wettlauf riskanter, weil man eventuell schneller rennt, ohne die Strecke zu kennen.

Globale, gemeinsame Regeln sollen helfen, KI-Risiken einzudämmen.

Zum Abschluss lohnt ein Blick auf die Bemühungen, den Wettlauf international zu zügeln. Im Oktober/November 2023 fanden erste *KI-Sicherheitsgipfel* statt (z. B. in Bletchley Park, United Kingdom), wo Vertreter aus den USA, der EU und China zusammentrafen, um Grundsätze zu erarbeiten. Allein dass China teilnahm, zeigt, dass man sich eines unregulierten Wettrüstens bewusst ist und Interesse an gewissen Absprachen haben könnte. Konkrete Ergebnisse waren etwa die *Bletchley Declaration*[56], in der gemeinsame Sorgen vor KI-Missbrauch geteilt und Folgetreffen vereinbart worden sind. Denkbar ist, dass große Modelle künftig meldepflichtig werden oder dass ein gewisser *Informationsaustausch über Sicherheitsmethoden* etabliert wird, selbst zwischen konkurrierenden Unternehmen, weil alle ein Interesse daran haben, dass eine fortgeschrittene KI, und sei es auch „nur" die eines Konkurrenten, außer Kontrolle gerät. Auf dem *AI Action Summit* in Paris kamen im Februar 2025 Vertreter und Vertreterinnen aus über 100 Ländern zusammen, um über die wirtschaftlichen Chancen und Herausforderungen von KI und regulatorische Standards zu diskutieren. Frankreichs Präsident Emmanuel Macron betonte die Bedeutung von KI für verschiedene Sektoren und rief zu internationaler Zusammenarbeit auf. In gewisser Weise sitzen alle Akteure im selben Boot: Ein wirklich katastrophaler KI-Unfall (ein KI-System verursacht großen realen Schaden, wirtschaftlich oder Personenschäden) könnte das Vertrauen in die gesamte Branche erschüttern. Daher besteht Hoffnung, dass trotz des Wettlaufs alle Akteure ein Minimum an vorsorglichen Sicherheitsmaßnahmen aufrechterhalten wird. Zusammengenommen ist der „Wettlauf zur AGI" Realität – er treibt die Technologie schneller voran, als viele erwartet hätten [Zweck und Werner 2025]. Die Teilnehmer – ob OpenAI, DeepMind, DeepSeek, Alibaba, Meta, Anthropic oder die Open-Source-Community – konkurrieren teils auf unterschiedliche Weise. Die entscheidende Frage aus Alignment-Sicht ist: *Wird dieser Wettbewerb fair und sicher ausgetragen oder endet er in einem ungebremsten Rennen, bei dem am Ende zwar jemand als Erster durchs Ziel fährt, aber alle die Bremsen verloren haben?* Genau deshalb plädieren viele Fachleute für Kooperation trotz Konkurrenz: Ein gewisses Maß an Absprachen, Sicherheitsstandards und ggf. Regulierung, damit der Weg zur AGI – wie ambitioniert er auch beschritten wird – nicht zur Gefahr für alle wird. Die nächsten Jahre (oder Monate?) werden zeigen, ob die Balance gelingt, sowohl *Fortschritt* als auch *Sicherheit* im Auge zu behalten, wenn wir uns der Verwirklichung immer autonomerer KI-Systeme nähern.

[56] The Bletchley Declaration by Countries, siehe https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, abgerufen am 25.02.2025

12 Fazit

AI-Alignment ist vom Randthema zu einer zentralen Herausforderung avanciert, während KI-Systeme eine immer größere Rolle in unserem Alltag und unserer Zukunftsplanung einnehmen. Die hier präsentierten Aspekte – von Definition und Notwendigkeit über Risiken und Beispiele bis hin zu Lösungen, Expertenstimmen und geopolitischen Dynamiken – zeigen ein facettenreiches Bild. Klar ist: Die Ausrichtung von KI auf menschliche Werte ist kein Selbstläufer, sondern erfordert bewusste Anstrengungen von Forschung, Entwicklung, Wirtschaft, Politik und der Gesellschaft. Die kommenden Entwicklungen in der KI werden bedeutend sein – aber es liegt an uns, dafür zu sorgen, dass sie kontrolliert und zum Wohle aller verlaufen. Die bevorstehenden Aufgaben sind anspruchsvoll, doch mit Umsicht, Kooperation und Innovation kann es gelingen, eine der entscheidendsten Technologien unserer Zeit auf Kurs zu bringen.

Literaturverzeichnis

Anthropic (2022), „Constitutional AI: Harmlessness from AI Feedback“, https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback, abgerufen am: 25.02.2025

Anthropic (2023), „Claude’s Constitution“, https://www.anthropic.com/news/claudes-constitution, abgerufen am 25.02.2025

Bengio, Y. (2024), „Reasoning through arguments against taking AI safety seriously“, https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/, abgerufen am: 25.02.2025

Booth, H. (2025), „When AI Thinks it will lose, it sometimes cheats, Study finds“, https://time.com/7259395/ai-chess-cheating-palisade-research/, abgerufen am 25.02.2025

Bubeck, S. et al. (2023), „Sparks of Artificial General Intelligence: Early experiments with GPT-4“, https://arxiv.org/abs/2303.12712

DeepSeek AI (2025), „DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning“, https://arxiv.org/pdf/2501.12948v1

Gawdat, M. (2022), „Scary Smart: Die Zukunft der künstlichen Intelligenz und wie wir mit ihrer Hilfe unseren Planeten retten“, Redline Verlag

Google Deepmind (2022), „Building safer dialogue Agents“, https://deepmind.google/discover/blog/building-safer-dialogue-agents/, abgerufen am 25.02.2025

Google (2022), „KI – zum Wohle der Gesellschaft“, https://about.google/intl/de_ZZ/stories/katharina-zweig/, abgerufen am 25.02.2025

Hastings-Woodhouse, S. (2024), „Introduction to Mechanistic Interpretability“, AI Safety Fundamentals, https://aisafetyfundamentals.com/blog/introduction-to-mechanistic-interpretability/, abgerufen am 25.02.2025

Medium, Operation Echo (2024), „Open-Source AI is uniquely dangerous“, https://medium.com/@operationecho/open-source-ai-is-uniquely-dangerous-0a3426cf0f40, abgerufen am 25.02.2025

Meinke, A. et al. (2024), „Frontier Models are capable of In-context Scheming“, Apollo Research, https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6751eb240ed3821a0161b45b/1733421863119/in_context_scheming_reasoning_paper.pdf, abgerufen am 25.02.2025

Mims, C. (2024), „This AI Pioneer thinks AI is dumber than a Cat“, https://www.wsj.com/tech/ai/yann-lecun-ai-meta-aa59e2f5, abgerufen am 25.02.2025

OpenAI (2023), „Introducing Superalignment – OpenAI Safety Research Report“, https://openai.com/index/introducing-superalignment/, abgerufen am 25.02.2025

OpenAI (2024), „GPT-4 System Card“, https://cdn.openai.com/papers/gpt-4-system-card.pdf, abgerufen am 25.02.2025

O'Connor, R., 2023, „Emergent Abilities of Large Language Models", https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models, abgerufen am 25.02.2025

Pan, X. et al. (2024), „Frontier AI Systems have surpassed the self-replicating red Line", https://arxiv.org/abs/2412.12140

Perrigo, B. (2023), „The new AI-Powered Bing is threatening Users. That's no laughing Matter", https://time.com/6256529/bing-openai-chatgpt-danger-alignment/, abgerufen am 25.02.2025

Roose, K. (2023), „A Conversation with Bing's Chatbot left me deeply unsettled", https://philosophy.tamucc.edu/texts/chat-with-chatgpt, abgerufen am 25.02.2025

Russell, S. (2022), „Of Myths and Moonshine". https://www.edge.org/conversation/the-myth-of-ai, abgerufen am 25.02.2025

Transformer Circuits (2022), „Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases", https://www.transformer-circuits.pub/2022/mech-interp-essay, abgerufen am 25.02.2025

Schneier, B. (2025), „AI Mistakes are very different from human Mistakes – We need new security systems designed to deal with their weirdness", https://spectrum.ieee.org/ai-mistakes-schneier, abgerufen am 25.02.2025

Wikipedia (2025), „AI Alignment – Grundlagen zur Ausrichtung von KI-Systemen und den damit verbundenen Herausforderungen", https://en.wikipedia.org/wiki/AI_alignment, abgerufen am 25.02.2025

World Economic Forum (2024a), „AI Value Alignment: Guiding Artificial Intelligence towards shared human Goals", https://www3.weforum.org/docs/WEF_AI_Value_Alignment_2024.pdf, abgerufen am 25.02.2025

World Economic Forum (2024b), „What is open-source AI and how could DeepSeek change the Industry?", https://www.weforum.org/stories/2025/02/open-source-ai-innovation-deepseek/, abgerufen am 25.02.2025

Xudong, P. et al. (2024), „Frontier AI systems have surpassed the self-replicating red Line", https://arxiv.org/abs/2412.12140

Zweck, A. und Werner, T. (2025, in Vorbereitung), Künstliche Intelligenz: Die Dynamik einer Technologie an der Schwelle zur Selbstoptimierung, VDI-Research

VDI Technologiezentrum GmbH
VDI Research
Airport City
VDI-Platz 1
40468 Düsseldorf

Telefon: +49 211 6214-536
E-Mail: foresight@vdi.de
www.vditz.de