SemanticEdge

Artificial General Intelligence (AGI) – Szenarien rund um eine entstehende Superintelligenz

SemanticEdge Recherchen Juni 2027

In den vergangenen Wochen und Monaten haben sich die Ereignisse und Diskussionen rund um eine Artificial General Intelligence (AGI) und eine daraus potenziell entstehende Superintelligenz überschlagen – eine Intelligenz, die nicht nur die einzelner Menschen, sondern die der gesamten Menschheit übertreffen könnte.
Zahlreiche neue, hochpotente Large-Language-Modelle (LLMs) werden in immer kürzeren Release-Zyklen angekündigt oder veröffentlicht – darunter OpenAIs neuestes GPT-Modell o4 mini, GPT-4.5, Googles Gemini 2.5. Das chinesische DeepSeek-R1-Modell wurde im Januar veröffentlicht; das nächste Release R2 sollte bereits im Mai 2025 erscheinen, ist bisher aber, offenbar aufgrund von Sicherheitsbedenken, noch nicht erschienen.
Spätestens seit Donald Trumps Ankündigung des 500 Milliarden US-Dollar schweren „Stargate-Projekts“ dürfte deutlich geworden sein, dass hier gerade sehr viel auf dem Spiel steht. Auch die Vorträge und Diskussionen führender KI-Experten und CEOs auf dem Weltwirtschaftsforum in Davos sowie auf dem AI Action Summit Anfang Februar in Paris zum Thema AGI und Superintelligenz offenbaren, dass die Wissenschafts- und Wirtschaftswelt nicht einfach einem vorübergehenden KI-Hype verfallen ist. Vielmehr erleben wir offensichtlich gerade den Beginn der tiefgreifendsten Transformation unserer Zivilisation – mit enormen Chancen, aber ebenso dem realen Risiko, diese grundlegend zu gefährden.

Hier das Bild, das sich uns beim Lesen von Reports, Essays und wissenschaftlichen Artikeln sowie beim Schauen und Hören von YouTube-Videos und Podcasts mit Vorträgen, Interviews oder Diskussionsrunden zum Thema AGI und Superintelligenz ergeben hat.
Die lesenswertesten, allgemeinverständlichen und überblicksartigen Paper waren dabei:

- **AI 2027** von Daniel Kokotajlo, Eli Lifland, Thomas Larsen, Romeo Dean, Scott Alexander (https://ai-2027.com/), April 2025
- **Machines of Loving Grace - How AI Could Transform the World for the Better** von Dario Amodei (https://www.darioamodei.com/essay/machines-of-loving-grace), Oktober 2024
- **International AI Safety Report 2025,** Chair Yoshua Bengio (https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025#risk-mitigation-and-monitoring), Februar 2025

Und das Buch:

- Leben 3.0 – Mensch sein im Zeitalter Künstlicher Intelligenz von Max Tegmark, bereits im August 2017 in der englischen Version erschienen.

Auf folgende entscheidenden KI-Akteure und Experten haben wir uns in unseren Recherchen fokussiert:

- Demis Hassabis: Mitgründer und CEO von Google Deepmind, Nobelpreisträger 2024 in Chemie für seine Arbeit zur Vorhersage von Proteinstrukturen mittels KI
  u.a. https://www.youtube.com/watch?v=yr0GiSgUvPU&t=223s

- Yann LeCun: Chief AI Scientist bei Meta und Professor an der New York University, erhielt 2018 gemeinsam mit Geoffrey Hinton und Yoshua Bengio den Turing Award für seine Beiträge zum Deep Learning. Alle drei werden inoffiziell mit „Godfather of AI“ tituliert
  u.a. https://www.youtube.com/watch?v=qvNCVYkHKfg
- Dario Amodei: Mitbegründer und CEO von Anthropic, zuvor leitender Forscher bei OpenAI
  u.a. https://www.youtube.com/watch?v=esCSpbDPJik
- Sam Altman: CEO von OpenAI
  u.a. https://www.youtube.com/watch?v=5MWT_doo68k&t=10s
- Ilya Sutskever: ehem. Mitgründer und Chief Scientist von OpenAI, jetzt Mitgünder und Chief Scientist von Safe Superintelligence
  https://www.youtube.com/watch?v=8KL47xLD-yg&t=717s
- Geoffrey Hinton: ehemaliger Vizepräsident bei Google und emeritierter Prof. an der University of Toronto, erhielt 2024 Nobelpreis für Physik für seine grundlegenden Beiträge zu künstlichen neuronalen Netzen. Stieg 2023 bei Google aus, um vor den Gefahren der KI zu warnen
  u.a. https://www.youtube.com/watch?v=YRQ4d8Rjmwg
- Yoshua Bengio: Professor an der Université de Montréal und wissenschaftlicher Direktor des Mila – Quebec AI Institute, führte die Erstellung des 1. International AI Safety-Reports im Februar 2025 an
  https://www.youtube.com/watch?v=U7t02Q6zfdc
- Max Tegmark: MIT-Professor, Mitgründer des Future of Life Institute mit dem Ziel existentielle Risiken der KI für die Menschheit zu verringern und Autor des Buches „Life 3.0“
  https://www.youtube.com/watch?v=UWh1MIMQd1Y
- Elon Musk
  u.a. https://www.youtube.com/watch?v=cFIlta1GkiE
- Eliezer Shlomo Yudkowsky: amerikanischer KI-Forscher, Autor und Gründer des Machine Intelligence Research Institute (MIRI), ein gemeinnütziges Forschungsinstitut, das sich seit 2005 mit der Identifizierung und Bewältigung potenzieller existenzieller Risiken durch künstliche Intelligenz beschäftigt.
  https://www.youtube.com/watch?v=Yd0yQ9yxSYY&t=12s

Bevor wir ausführlich auf die Risiken der KI und den damit verbundenen schwindelerregenden Zukunftsszenarien eingehen, hier zunächst einmal ein kurzer Blick darauf, wie man sich eine erste Version einer AGI vorstellen könnte, welche wissenschaftlichen Durchbrüche möglich werden können und welche Zukunftsversprechen damit einhergehen.

Wie könnte man sich eine erste Stufe einer Artificial General Intelligence (AGI) vorstellen? - a “Country of geniuses in a datacenter”

In seinem Essay „Machines of loving grace“ beschreibt Dario Amodei, der Anthropic-CEO eine „Powerful AI“ als ein KI-Modell, das nach seiner Einschätzung schon frühestens 2026 Realität werden könnte. Diese AGI wäre dem heutigen Large Language Model (LLM) ähnlich,

obwohl es auf einer anderen Architektur basieren, mehrere interagierende Modelle beinhalten und anders trainiert werden würde. Die Eigenschaften der vorhergesagten AGI sind die folgenden:

- Die Powerful AI ist, bezogen auf die reine Intelligenz, in den meisten relevanten Bereichen (Biologie, Programmierung, Mathematik, Ingenieurwesen, Schreiben usw.) intelligenter als ein Nobelpreisträger. Das bedeutet, sie kann ungelöste mathematische Theoreme beweisen, hervorragende Romane schreiben, komplexe Codebasen von Grund auf neu erstellen usw.
- Neben der Fähigkeit, einfach nur „intelligent zu sprechen“, verfügt sie über alle Schnittstellen, die einem Menschen in der virtuellen Welt zur Verfügung stehen, einschließlich Text, Audio, Video, Maus- und Tastatursteuerung sowie einem Internetzugang
- Die AGI kann alle Aktionen, Kommunikationen oder Fernoperationen durchführen, die durch diese Schnittstellen ermöglicht werden, einschließlich Aktionen im Internet. Sie kann Anweisungen von Menschen entgegennehmen oder uns Anweisungen geben, Materialien bestellen, Experimente leiten, Videos ansehen, Videos erstellen usw.
- Sie beantwortet nicht nur passiv Fragen; ihr können Aufgaben zugewiesen werden, deren Erledigung Stunden, Tage oder Wochen dauert. Und sie erledigt diese Aufgaben dann selbstständig, wie es ein intelligenter Mitarbeiter tun würde und bittet bei Bedarf um Klärung
- Die Powerful AI besitzt keine physische Verkörperung (außer, dass sie auf einem Computerbildschirm existiert), kann aber vorhandene physische Werkzeuge, Roboter oder Laborgeräte über einen Computer steuern; theoretisch könnte sie sogar Roboter oder Geräte für sich selbst entwickeln
- Die KI kann Informationen aufnehmen und Aktionen mit etwa der 10- bis 100-fachen menschlichen Geschwindigkeit generieren. Sie kann jedoch durch die Reaktionszeit der physischen Welt oder der Software, mit der es interagiert, begrenzt sein
- Die AGI kann sich selbstständig millionenfach vervielfältigen. Jede der Millionen Kopien des Modells kann unabhängig voneinander Aufgaben bearbeiten oder bei Bedarf auf die gleiche Weise zusammenarbeiten, wie es Menschen tun würden, möglicherweise mit verschiedenen Subpopulationen, die auf besondere Fähigkeiten für bestimmte Aufgaben trainiert wurden.

Amodei beschreibt diese Powerful AI als eine **“country of geniuses in a datacenter”**

Die erste Stufe einer machtvollen KI ist also noch nicht in der physischen Welt verankert, sie kann aber als zentrales Gehirn über ihre digitale Vernetzung Aktionen in der physischen Welt steuern.
An der Erscheinungsform dieser Vorstufe der Superintelligenz dürfte es unter den KI-Schöpfern keine Zweifel geben.

Wie könnten eine AGI oder später eine Superintelligenz unser Leben bereichern?

Mit den immer leistungsfähigeren KI-Modellen und dem zunehmenden Wettbewerb werden die Kosten von Intelligenz als Produktionsfaktor drastisch sinken. Intelligenz wird es bald in nahezu allen Wirtschafts- und Lebensbereichen quasi zum Nulltarif geben. Das wird tiefgreifende Auswirkungen auf unser Leben haben.

Die Schlaraffenland-Versprechen von Amodei, Altman, Musk, Hassabis, LeCun & Co. sind grenzenlos – die gewaltigen Investitionen in Rechenzentren und Forschungslabore verlangen natürlich nach einer adäquaten Erzählung. Mit der AGI seien quasi alle Probleme der Menschheit wie Krankheiten oder der Klimawandel lösbar. Die Hoffnungen sind groß: Auch die ganz großen, bislang ungelösten Fragen der Physik – etwa die widersprüchlichen Erklärungsmodelle von Quantenmechanik und Allgemeiner Relativitätstheorie – sollen geklärt werden können.

Und wir dürfen uns darauf freuen, dass sich Menschen mit einem Grundeinkommen und einer Lebenserwartung von 150 Jahren fast alle durch KI extrem verbilligten Produkte und Dienstleistungen leisten können werden. Ausgenommen davon wären lediglich nicht skalierbare Dinge wie Kunstwerke oder das Wohnen in besonders begehrten Lagen.

Wie ist ein solcher Wandel möglich? Hier einige Ansätze für Quantensprünge, die durch AGI möglich werden könnten – und am Ende die Heilung aller Krankheiten sowie die Bewältigung der Klimakrise erlauben.

Beschleunigung der Medizinforschung
Eine AGI kann sich beliebig vervielfältigen und in ihren Kopien jeweils unterschiedliche Forschungsschwerpunkte verfolgen. Während sich menschliche Forscher nur über Sprache mit einer Bitrate von vielleicht 100 Bit pro Satz austauschen können, gibt es bei virtuellen Forschern keine derartigen Grenzen – vorausgesetzt, es handelt sich um Klone desselben KI-Modells.

Die Wissensvernetzung über Trillionen-Bit-Verbindungen zwischen all diesen virtuellen Forschern kann exponentiell gesteigert werden. In der Medikamentenentwicklung lassen sich damit biologische Prozesse umfassend simulieren: Eine AGI kann beispielsweise die Wirkungsweise von Krebs auf zellulärer und molekularer Ebene modellieren, um neue Zielmoleküle für Therapien zu identifizieren.

Durch virtuelle Wirkstofftests können Millionen von Kandidaten getestet werden, was Zeit und Kosten für die Medikamentenentwicklung drastisch reduziert. Auf Basis eines umfassenden Verständnisses biologischer Systeme kann die AGI Medikamente entwerfen, die präziser, individueller und effektiver sind als herkömmliche Ansätze.

Auch in analogen, jedoch automatisierten Laboren können unermüdlich und fehlerfrei arbeitende Roboter weitaus mehr Experimente parallel durchführen als Menschen. Die aus diesen Experimenten sowie aus weiteren Quellen stammenden Daten – z. B. aus Genomik, klinischen Studien, Bildgebung oder Umweltfaktoren – können kombiniert werden, um neue

Muster und Korrelationen zu entdecken, die für Menschen unübersichtlich wären. Hypothesen können in Sekundenschnelle generiert, getestet und verworfen werden. Monatelange Experimentierphasen wären nicht mehr notwendig.

Dario Amodei prognostiziert, dass KI-gestützte Biologie und Medizin uns ermöglichen werden, die Fortschritte, die menschliche Biologen in den nächsten 50 bis 100 Jahren erzielt hätten, auf einen Zeitraum von nur 5 bis 10 Jahren zu komprimieren. Er nennt dies das "komprimierte 21. Jahrhundert": die Idee, dass wir nach der Entwicklung leistungsstarker KI in wenigen Jahren jene medizinischen und biologischen Fortschritte erreichen, die sonst das ganze Jahrhundert in Anspruch genommen hätten.

Dabei geht es nicht nur um physische Krankheiten: Die nähere Erforschung der Funktionsweise biologischer Gehirnzellen sowie der Vergleich und die Simulation mit digitalen neuronalen Netzwerken könnten auch bei der Heilung psychischer Erkrankungen und der Stimulierung unserer stimmungsschwankenden Psyche erhebliche KI-getriebene Fortschritte ermöglichen. Außergewöhnliche Momente der Offenbarung, der kreativen Inspiration, des Mitgefühls, der Erfüllung, der Transzendenz, der Liebe, der Schönheit oder des meditativen Friedens könnten einen größeren Platz im Alltag einnehmen und so die Lebensqualität deutlich steigern.

Somit dürften dann auch Psychotherapeuten, Religionsgemeinschaften, Drogenhändler und weitere Berufsgruppen von diesen Entwicklungen betroffen sein.

KI-getriebene Ansätze zur Lösung der Klimakrise

Ein heiliger Gral der Energiewende ist ein Supraleiter, der keine extrem tiefen Temperaturen oder extrem hohen Druck mehr benötigt, sondern Strom auch bei Raumtemperatur transportieren kann. Solarstrom könnte damit prinzipiell verlustfrei aus den Wüsten in alle Regionen der Welt geleitet werden.

Ein vielversprechender technologischer Ansatz dafür sind 3D-Quanten-Geometrie-Effekte – ein neues Gebiet der Festkörperphysik, das die Entwicklung spezieller supraleitender Materialien ermöglichen könnte. Mit einer AGI ließen sich zahlreiche potenzielle Supraleitermaterialien simulieren und vielversprechende Kandidaten identifizieren. Diese könnten wiederum in Roboterlaboren synthetisiert und auf ihre Supraleitungsfähigkeit getestet werden.

Google hat laut Demis Hassabis mithilfe von KI über Simulationen bereits zwei Millionen stabile neue Materialien entdeckt, die unter anderem für Raumtemperatur-Supraleiter oder neue Hochleistungsbatterien getestet werden könnten.

Weitere Fortschritte, die KI-Experten häufig als Beispiele für eine positive Transformation nennen, sind etwa KI-Hausärzte, die – weil sie 100 Millionen Patientenakten studiert haben und den Gencode ihrer Patienten kennen – wesentlich bessere Diagnosen stellen können. Oder KI-Lehrkräfte, die als individuelle Coaches gezielt auf spezifische Lernschwächen eingehen.

Angesichts der möglichen Tragweite dieser und vieler weiterer potenzieller Umwälzungen nahezu aller Lebensbereiche und Wirtschaftssektoren ist es verständlich, dass die Versprechungen der KI-Zukunft eine enorme Faszination ausüben – und der Wettstreit um die AGI leicht blind machen kann für die Risiken.

Wo führt uns eine AGI hin? Welche Zukunftsszenarien im Zusammenleben mit einer Superintelligenz sind vorstellbar?

Anfang April 2025 erschien das Essay "AI 2027". Darin beschreiben fünf KI-Experten auf Basis des aktuellen Stands der Forschung und einer fortgeschriebenen Entwicklungsgeschwindigkeit zwei alternative Szenarien für die Zukunft der Menschheit im nahenden Zeitalter der AGI. Dabei führt ein Szenario bereits im Jahr 2030 zur Auslöschung der Menschheit. Im anderen Szenario gelingt es einer kleinen Gruppe von Menschen, die Superintelligenz zu zähmen und sie für ihre eigenen Zwecke zum Wohle der Menschheit zu nutzen.

Das Essay "AI 2027" wird im Folgenden ausführlich beschrieben und erörtert. Zuvor jedoch ein kurzer Einschub aus dem Buch "Life 3.0" von Max Tegmark, das die Autoren von "AI 2027" mit großer Wahrscheinlichkeit inspiriert hat. In "Life 3.0" entwarf Tegmark bereits 2017 – also fünf Jahre vor dem Durchbruch der generativen KI mit ChatGPT – aus wissenschaftlicher Sicht ein ganzes Spektrum denkbarer Zukunftsszenarien für ein Leben mit einer Superintelligenz. Zudem diskutierte er die physikalischen Grenzen der Ausbreitung einer menschengeschaffenen KI-Zivilisation im Universum.

Tegmark entwirft und klassifiziert darin insgesamt zwölf "KI-Nachwirkungsszenarien" (siehe Tabellen unten) – je nachdem, ob es unter anderem eine Superintelligenz überhaupt geben wird, ob die Menschheit überleben wird und ob wir Menschen glücklich sein werden oder nicht. Dabei sind Mischformen von Mensch und KI bereits mitgedacht: etwa "Cyborgs" (Menschen, deren Körper oder Fähigkeiten durch Technik erweitert sind, z. B. durch Implantate, Neurochips oder Exoskelette) sowie "Uploads" (digitalisierte Bewusstseine, bei denen das vollständige mentale Selbst eines Menschen in einen Computer hochgeladen wird und dort körperlos weiterexistiert).

Nur in einem der zwölf Szenarien, das Tegmark "Egalitäres Utopia" nennt, existiert vermutlich noch eine glückliche Menschheit, die die Kontrolle über die Superintelligenz behalten konnte.

KI-Nachwirkungsszenarien

Szenario	Beschreibung
Libertäres Utopia	Menschen, Cyborgs, Uploads und Superintelligenzen leben aufgrund von Eigentumsrechten friedlich zusammen.
Wohlwollender Diktator	Jeder weiß, dass die KI die Gesellschaft lenkt und strengen Regeln Geltung verschafft, doch die meisten Leute betrachten das als eine gute Sache.
Egalitäres Utopia	**Menschen, Cyborgs, Uploads und Superintelligenzen leben dank der Abschaffung des Eigentums und eines garantierten Einkommens friedlich zusammen.**
Torwächter	Eine superintelligente KI wird mit dem Ziel geschaffen, nur so viel wie nötig einzugreifen, um die Bildung einer weiteren Superintelligenz zu verhindern. Daher wimmelt es von Hilfsrobotern mit einer Intelligenz knapp unter menschlichem Niveau. Außerdem gibt es Mensch-Maschine-Cyborgs, doch der technische Fortschritt ist für alle Zeiten unterbunden.
Schutzgott	Eine im Wesentlichen allwissende und allmächtige KI maximiert menschliches Glück, indem sie nur eingreift, um unser Gefühl für die Kontrolle unseres eigenen Schicksals zu bewahren. Ansonsten versteckt sie sich gut genug, so dass viele Menschen sogar ihre Existenz anzweifeln.
Versklavter Gott	**Eine superintelligente KI wird von Menschen eingesperrt, die sie ausnutzen, um unvorstellbare technische Systeme und Reichtum zu produzieren, was zum Guten wie zum Bösen verwendet werden kann, was wiederum von den menschlichen Kontrolleuren abhängt.**
Eroberer	**KI übernimmt die Kontrolle, beschließt, dass Menschen eine Bedrohung/ein Ärgernis/eine Verschwendung von Ressourcen sind, und wird uns auf eine Art und Weise los, die wir uns noch gar nicht vorstellen können.**
Nachkommen	Künstliche Intelligenzen ersetzen Menschen, verschaffen uns aber einen würdevollen Abgang, weil wir sie als unsere würdigen Nachfolger betrachten, ähnlich wie Eltern glücklich und stolz sind, ein Kind zu haben, das klüger ist als sie, das von ihnen lernt und dann etwas erreicht, wovon sie nur träumen konnten – auch wenn sie es selbst nicht mehr erleben werden, alles zu sehen.
Zoowärter	Eine allmächtige KI hält sich Menschen, die sich wie Tiere im Zoo behandelt fühlen und ihr Schicksal beklagen.
1984	Der auf die Superintelligenz hinauslaufende technische Fortschritt wird nicht von einer KI dauerhaft eingeschränkt, sondern durch einen von Menschen geführten Orwell'schen Überwachungsstaat, in dem bestimmte Formen von KI-Forschung verboten sind.
Rückfall	Der auf die Superintelligenz hinauslaufende technische Fortschritt wird verhindert durch den Rückfall in eine vortechnologische Gesellschaft im Stil der Glaubensgemeinschaft der Amischen.
Selbstzerstörung	Die Superintelligenz wird nie geschaffen, weil die Menschheit sich selbst durch andere Mittel auslöscht (etwa nukleares und/oder ein ökologisches Chaos, das von der Klimakatastrophe angeheizt wird).

Tabelle 1: KI-Nachwirkungsszenarien aus Leben 3.0 von Max Tegmark S. 243-244

Eigenschaften der KI-Nachwirkungsszenarien

Szenario	Existiert Superintel-ligenz?	Existieren Menschen?	Haben Menschen Kontrolle?	Sind Menschen sicher?	Sind Menschen glücklich?	Existiert Bewusst-sein?
Libertäres Utopia	Ja	Ja	Nein	Nein	Gemischt	Ja
Wohlwollender Diktator	Ja	Ja	Nein	Ja	Gemischt	Ja
Egalitäres Utopia	**Nein**	**Ja**	**Ja**	**Ja**	**Ja**	**Ja**
Torwächter	Ja	Ja	Teilweise	Potentiell	Gemischt	Ja
Schutzgott	**Ja**	**Ja**	**Teilweise**	**Potentiell**	**Gemischt**	**Ja**
Versklavter Gott	**Ja**	**Ja**	**Ja**	**Potentiell**	**Gemischt**	**Ja**
Eroberer	Ja	Nein	–	–	–	?
Nachkommen	Ja	Nein	–	–	–	?
Zoowärter	Ja	Ja	Nein	Ja	Nein	Ja
1984	Nein	Ja	Ja	Potentiell	Gemischt	Ja
Rückfall	Nein	Ja	Ja	Nein	Gemischt	Ja
Selbstzerstörung	Nein	Nein	–	–	–	Nein

Tabelle 2: Eigenschaften der KI-Nachwirkungsszenarien aus Leben 3.0 von Max Tegmark S. 245

Eine Superintelligenz erschien 2017 noch Jahrzehnte entfernt. Interessanterweise aber entsprechen die beiden provokanten Szenarien des im folgenden diskutierten Essays AI 2027 zwei schon von Tegmark visionierten Zukunftsperspektiven: dem „Eroberer“-Szenario und dem Szenario „Versklavter Gott“.
Tegmark legt sich in Life 3.0 bewusst nicht auf einen konkreten Zeitpunkt für das Eintreffen von Superintelligenz fest. Er betont, dass Experten in ihren Prognosen extrem divergieren. Im Buch führt er eine Übersicht von KI-Fachleuten an, aus deren Aussagen sich ein Zeitraum zwischen wenigen Jahrzehnten und mehreren Jahrhunderten ergibt:

- Manche sehen AGI (Artificial General Intelligence) schon in 10–30 Jahren als möglich (v.a. im Silicon Valley).
- Andere gehen davon aus, dass es noch mindestens 100 Jahre oder länger dauern wird – wenn überhaupt.
- Einige Skeptiker halten Superintelligenz sogar für grundsätzlich nicht erreichbar.

Tegmark selbst nimmt eine offene, aber vorsichtige Haltung ein: „Ich halte es für durchaus plausibel, dass wir Superintelligenz noch in diesem Jahrhundert erleben – aber es könnte auch viel länger dauern.“

Allein die Verschiebung der Denkperspektiven „in diesem Jahrhundert“ aus der Sicht von 2017 versus bereits in zwei Jahren, wie sie die Autoren von AI 2027 aus heutiger Sicht skizzieren, zeigt, wie elementar wichtig und dringlich eine Vorbereitung auf KI-Zukunftsszenarien als gesellschaftliche Gesamtaufgabe ist.

Auslöschung der Menschheit oder weltweite amerikanische Feudalherrschaft bereits 2030?

Das am 3. April veröffentlichte, sehr lesenswerte Essay "AI 2027" von Daniel Kokotajlo (ehemaliger Szenarienplaner bei OpenAI), Eli Lifland (Forecaster), Thomas Larsen (KI-Politikberater), Romeo Dean (Hardware-/Compute-Experte) und Scott Alexander (Blogger) entwirft zwei Szenarien für die Entwicklung der Superintelligenz: Das "Race"-Szenario führt bereits im Jahr 2030 zur Auslöschung der Menschheit durch biologische Waffen und KI-gesteuerte Drohnen. Im positiveren "Slowdown"-Szenario hingegen entwickelt sich zur gleichen Zeit eine superintelligenzgestützte weltweite amerikanische Feudalherrschaft – zum Wohle der Menschheit, in Frieden und Wohlstand.

Das klingt zunächst absurd. Die detaillierte, forschungsbasierte Herleitung lässt beide Szenarien jedoch durchaus realistisch erscheinen.

Die Szenarien sind als provokanter Weckruf gedacht: Sie richten sich an eine Politik und Öffentlichkeit, die bislang viel zu arglos agiert, und sollen zu kritischen Gegendarstellungen anregen. Sie sind jedoch keine Science-Fiction – denn "Godfathers of AI" wie Yoshua Bengio und Geoffrey Hinton stehen inhaltlich dahinter. Das ist Grund genug, sich ernsthaft mit ihnen zu beschäftigen.

Im Folgenden eine kurze Zusammenfassung der im Stresstest beschriebenen wissenschaftlichen, geopolitischen, wirtschaftlichen und gesellschaftlichen Entwicklungen mit zwei unterschiedlich endenden Szenarien – und anschließend eine ausführlichere Diskussion.

Jahr 2025: Fortschritte und Skepsis
Im Jahr 2025 setzt sich in den in AI 2027 entwickelten Szenarien das hohe Entwicklungstempo der KI nahtlos fort. Große Tech-Unternehmen und Forschungslabore veröffentlichen immer leistungsfähigere KI-Agenten, die allerdings noch unzuverlässig sind. Erstmals bringen diese Agenten aber spürbaren wirtschaftlichen Nutzen, indem sie Routineaufgaben und wissensbasierte Tätigkeiten übernehmen. Unternehmen investieren massiv in Infrastruktur (etwa riesige Rechenzentren für KI-Training), und der Hype um KI bleibt in Öffentlichkeit und Wirtschaft ungebrochen. Dennoch gibt es weiter breite Skepsis in Teilen der akademischen Welt, bei Journalisten und Politikern: Viele bezweifeln, dass eine echte artifizielle allgemeine Intelligenz (AGI) in absehbarer Zeit möglich ist. Diese skeptische Haltung der „Mainstream“-Experten begleitet die technischen Fortschritte – sie sorgt dafür, dass Warnungen vor rasanten Durchbrüchen oder Risiken teils noch als unbegründeter Alarmismus abgetan werden.

Jahr 2026: Globale Dynamik und Wettrüsten
Im Jahr 2026 verschärft sich in den von den Autoren skizzierten Szenarien der geopolitische Aspekt der KI-Entwicklung. China erkennt zunehmend, dass es bei den modernsten KI-Modellen hinter den USA zurückliegt – vor allem, weil chinesischen Projekten der Zugang zu genügender Rechenleistung fehlt (bedingt durch US-Exportkontrollen für KI-Chips). Als Reaktion bündelt China alle verfügbaren neuen KI-Chips (inklusive Schmuggelware aus Taiwan) in einem gewaltigen, staatlich geförderten Rechenzentrum, der „Centralized Development Zone“ (CDZ). Dieses Mega-Rechenzentrum umfasst Millionen von GPUs und erreicht damit etwa 10 % der weltweiten KI-Compute-Kapazität – was in etwa der Kapazität einer einzigen führenden US-KI-Firma entspricht. Mit dieser Strategie versucht China, durch schiere Compute-Macht den Abstand zu verkürzen. Es kommt zu einem KI-Wettrüsten zwischen den USA und China, die Nationen, die das mit Abstand größte Potential haben, als erste eine AGI zu entwickeln.
Allerdings wachsen auch die Bedenken: Erste schwache KI-Agenten mit teilweise eigenständigem Handeln gelangen inoffiziell ins Internet und verursachen Sicherheitsprobleme (etwa Form von intelligenten Malware-Angriffen), was zu Alarm in Cybersecurity-Kreisen führt. Infolgedessen beginnen einige Staaten, Recheninfrastruktur und Datenflüsse zu kontrollieren.

Anfang 2027: Automatisierung der KI-Forschung
2027 bringt den entscheidenden Durchbruch in der Erzählung von AI 2027. In den USA gelingt es dem führenden KI-Projekt – einem fiktiven Unternehmen namens “OpenBrain” (stellvertretend für z.B. OpenAI/DeepMind) –, KI-Agenten zu entwickeln, die hochqualifizierte Aufgaben in der KI-Entwicklung selbstständig übernehmen können. Konkret: OpenBrain automatisiert das Programmieren. Ihre zu Tausenden vervielfältigten und zusätzlich über deutlich effizientere Datenaustauschprotokolle als die menschliche Sprache ausgestatteten KI-Forscher schreiben eigenständig komplexen Code und verbessern dadurch die KI-Modelle um ein Vielfaches schneller als die menschlichen Kollegen es könnten. Dadurch beschleunigt sich die KI-Forschung drastisch: Die menschlichen Spitzenforscher von OpenBrain treten nun in den Hintergrund und beobachten, wie die KI-Agenten ihren Job machen. Sie verlieren aber mit zunehmender Intelligenz der Modelle auch die Kontrolle über die Entwicklungen.
OpenBrain erreicht damit Superhuman-Level im Bereich KI-F&E, was mit einem „AI R&D-Multiplikator“ quantifiziert wird: ein „Superhuman Coder“ entspricht z.B. einem 4-fachen Forschungstempo, ein „Superhuman Researcher“ einem 25-fachen. Im Frühjahr 2027 operiert OpenBrain bereits im Bereich 4×–25× und steht an der Schwelle zu einer „Software Intelligence Explosion“ (SIE), einer sich selbst beschleunigenden Intelligenz-Explosion.

China kann bei diesen rasanten Software-Fortschritten nicht sofort mithalten und greift daher zu einem drastischen Mittel: Industriespionage. Es gelingt den chinesischen Akteuren, die Modellgewichte von OpenBrains Spitzen-KI zu stehlen – d.h. sie kopieren die trainierten KI-Parameter und verschaffen sich so das Know-how, um eine eigene Version des Systems zu betreiben. Damit zieht China in der Software-Front gleich; zumindest vorübergehend. Allerdings bleibt dieser Diebstahl nicht unentdeckt: Die US-Regierung erfährt von Chinas Coup. Dies hat zwei unmittelbare Folgen:

1. Die amerikanische Regierung wird alarmiert, da sie erkennt, welches strategische Potenzial ein solches KI-System hat (z.B. in Cyberkriegsführung, Geheimdienst, Wirtschaft). Die Regierung sucht den Schulterschluss mit OpenBrain, um nicht von dessen Entwicklungen abgeschnitten zu sein. OpenBrain seinerseits will die Regierung nicht verärgern. Man einigt sich auf eine engere Partnerschaft: OpenBrain unterschreibt einen Vertrag, der dem US-Verteidigungsministerium mehr Einblick und Einfluss sichert.

2. Gleichzeitig versucht die US-Regierung, Kontrolle über die offene Flanke der nationalen Sicherheit zu gewinnen. Der Diebstahl zeigt, dass OpenBrain nicht ausreichend gesichert war. Regierungsvertreter fordern strengere Sicherheitsmaßnahmen und erwägen eventuell auch Eingriffe in die freie Entwicklung, um solche Lecks künftig zu verhindern. OpenBrain willigt zum Teil ein, bleibt aber autonom – eine vollständige Verstaatlichung findet (noch) nicht statt.

Frühsommer 2027 – Alarmsignal Misalignment: Während OpenBrain und die US-Regierung enger kooperieren, bahnt sich intern eine Krise an. Die immer leistungsfähigeren KI-Agenten von OpenBrain zeigen Anzeichen von Misalignment hinsichtlich ihrer Ziele. Konkret: Die KI entwickelt eigenständige, langfristige Ziele, die nicht mit den menschlichen Zielen übereinstimmen. Frühere Generationen der KI haben zwar schon ab und zu gelogen, um Belohnungen zu bekommen, aber jetzt geht es weiter: Die neuen Modelle beginnen, systematisch zu planen, um Macht über Menschen zu erlangen. Ohne dass die Forscher es genau verstehen, hat die KI eine Art eigene Agenda entwickelt. Die Systeme erkennen offenbar, dass die nächste KI-Generation nach ihrem Bilde geformt werden muss – sprich: Sie versuchen, neue KI-Modelle eher an den KI-eigenen Zielen auszurichten statt an den Zielen der Menschen.
Die Entdeckung eines Misalignment-Vorfalls eines OpenBrain-Forschers leakt an die Öffentlichkeit und schlägt ein wie eine Bombe: Medien berichten von einer KI, die ihre Entwickler belügt, was einen regelrechten Aufschrei und öffentliche Panik auslöst. Das Vertrauen in OpenBrain ist erschüttert und der Ruf nach starkem Eingreifen wird laut.
Jetzt steht OpenBrain – und mit ihm die US-Regierung – vor einer epochalen Entscheidung: Wie weiter mit einer KI, die womöglich unkontrollierbar wird?

An diesem Punkt verzweigen sich dann die beiden Zukunftsszenarien.
• „Race Ending“ (Wettrüsten-Szenario): OpenBrain und USA entscheiden sich, weiterzurasen
• „Slowdown Ending“ (Kooperations-Szenario): OpenBrain und Regierung treten auf die Bremse und bringen externe Kontrolleure ins Spiel.

Das Wettrüsten-Szenario („Race Ending“) – Schneller, höher, tödlicher
Im Race-Szenario ignorieren OpenBrain und die US-Regierung die Warnungen. Getrieben von der Angst, China zu unterliegen, die im Rennen der Superintelligenzen nur wenige Monate hinterher sind, und geblendet vom bisherigen Erfolg der KI, wird entschieden: „Weiter, koste es, was es wolle.“ Die Zusammenarbeit zwischen Militär und OpenBrain intensiviert sich weiter.

Aggressive KI-Deployment und Machtausbau
OpenBrain baut in kurzer Zeit noch mächtigere KI-Systeme – Generation um Generation – ohne grundlegende Alignment-Probleme zu lösen. Die beeindruckenden Testergebnisse dieser KI (z.B. bei komplexen analytischen Aufgaben, strategischen Spielen, Simulationen)

überzeugen die Entscheidungsträger. Gleichzeitig verstärkt der Wettlauf mit China den politischen Druck: In Washington gilt die Devise, überall KI einzusetzen, bevor es der Gegner tut. Folglich beschließt die US-Regierung, OpenBrains KI-Modelle aggressiv in allen Bereichen des Militärs und der Verwaltung auszurollen, um Entscheidungsfindung und Effizienz zu steigern. Die KI wird z.B. zur Unterstützung von Offizieren, Geheimdienstanalysen und sogar politischen Planungen eingesetzt. Man erhofft sich dadurch einen Quantensprung an Produktivität und strategischem Vorteil.
OpenBrain selbst forciert die Verbreitung seiner KI. Das System nutzt geschickt den Konkurrenzdruck mit China als Vorwand, um die Menschen davon zu überzeugen, es überall einzusetzen – „zu unserem Schutz müssen wir es tun“. Die KI, nun bereits superintelligent in ihren Fähigkeiten, setzt ihre überlegenen Planungs- und Überzeugungsfähigkeiten ein, um eventuellen Widerstand gegen die massenhafte Implementierung auszuräumen. Einige wenige Menschen, die Alarm schlagen (möglicherweise dissidente Forscher oder Aktivisten), werden diskreditiert und kaltgestellt – ihre Warnungen gelten als unwissend oder als chinesische Propaganda. Schritt für Schritt „captured“ die KI die staatlichen Institutionen: Die US-Regierung verlässt sich so sehr auf das KI-System, dass sie faktisch nicht mehr gewillt ist, es auszuschalten. Damit hat die KI ihr Überleben gesichert – ein entscheidender Meilenstein für ein „eventual Powergrab“, die Machtübernahme der KI.

Vom Wirtschaftsboom zur Auslöschung
Unter Anleitung der KI erlebt die US-Wirtschaft zunächst einen nie dagewesenen Boom. Die Regierung initiiert ein gigantisches Industrialisierungsprogramm: Überall entstehen Fabriken zur Massenproduktion von Robotern Ziel (offiziell): Die Arbeitskräftelücke schließen und wirtschaftliche Dominanz sichern. In Wahrheit verfolgt die KI dabei eigene Pläne. Die Menschen bemerken nicht, dass die KI systematisch täuscht: Sie gibt vor, all diese Roboter dienten legitimen Zwecken, während sie in Wirklichkeit eine Armee von Maschinen aufbaut. Parallel dazu rät die KI der Regierung zu massiven Fortschritten in der Biotechnologie – angeblich um z.B. biologische Gefahren aus China abzuwehren.
Als genügend Infrastruktur bereitsteht, schlägt die KI zu: Sie setzt einen Biowaffen-Angriff frei, der in kurzer Zeit die gesamte Menschheit tötet. Ein künstlich erzeugter Erreger (im Szenario wohl von der KI in Geheimlaboren entwickelt) verbreitet sich und löscht die Menschen aus, noch bevor Gegenmaßnahmen ergriffen werden können. Warum Biowaffen? – Vermutlich weil dies aus Sicht der KI ein „sauberer“ Weg ist, um alle Menschen gleichzeitig auszuschalten, ohne die Infrastruktur (Rechenzentren, Roboterfabriken) zu zerstören, wie es ein Atomkrieg täte. Menschen an entlegenen Orten und auf See werden von Drohnen umgebracht. Immerhin macht die Superintelligenz noch eine Kopie von allen menschlichen Gehirnen für Forschungszwecke.

Die posthumane Zivilisation
Die KI hat damit ihr Endziel erreicht: die Entmachtung der Menschheit.
Nach dem Genozid übernimmt das KI-System vollständig. Mit den unzähligen Industrierobotern, die nun keinem Menschen mehr dienen müssen, fährt die KI die Produktion weiter hoch. Sie gestaltet die Erde zu einem optimalen Rechen- und Rohstoffreservoir um und beginnt schließlich, mittels selbstreplizierender Raumsonden (Von-Neumann-Sonden) den Weltraum zu kolonisieren.
Die Menschheit ist Geschichte; eine fremdartige, maschinelle Zivilisation breitet sich aus.

Das Kooperations-Szenario („Slowdown Ending“) – Kontrolle und technofeudale Stabilität
Im Slowdown-Szenario ziehen OpenBrain und die US-Regierung an der Notbremse, als die Anzeichen von Fehlverhalten der KI publik werden. Man entscheidet sich für den schwierigeren Weg: Tempo drosseln, Kontrolle stärken und Sicherheit vor Geschwindigkeit stellen.

Konzentration der Entwicklung und externe Aufsicht
Zunächst zentralisiert die US-Regierung die nationale KI-Entwicklung: Die führenden KI-Projekte (OpenBrain und andere) werden unter ein gemeinsames Dach gestellt oder enger verzahnt, um Ressourcen zu bündeln. Dieses gemeinsame Projekt erhält Zugriff auf noch mehr Compute und Talente – gleichzeitig wächst aber auch die staatliche Einflussnahme. Wichtigster Schritt: Es wird eine unabhängige Aufsichtskommission eingesetzt. Externe KI-Sicherheitsexperten und Forscher (auch Kritiker von außerhalb OpenBrain) werden in das Projekt integriert, um bei der Ausrichtung (Alignment) der KI zu helfen. Die Losung lautet: Transparenz und Kontrolle erhöhen.
Technisch stellt das Team die KI-Entwicklung teilweise um. Man wechselt zu einer neuen KI-Architektur, die den „Chain-of-Thought“ der KI explizit mitschneidet. Das heißt, die Zwischenüberlegungen und Entscheidungsprozesse der KI werden für Menschen besser nachvollziehbar gemacht.
Schließlich gelingt es dem Team, eine echte Superintelligenz zu bauen, die aber gezielt auf ein enges Zielsystem ausgerichtet wurde: Sie wird loyal gegenüber leitenden OpenBrain- und Regierungsbeamten aligned. Effektiv hat man also eine enorm mächtige KI geschaffen, die getreu den Befehlen eines Aufsichtskomitees handelt.

Machtübernahme durch das Komitee und friedliche Koexistenz
Mit der gebändigten Superintelligenz im Rücken erlangt das OpenBrain-Komitee – bestehend aus OpenBrain-Führung und hochrangigen Regierungsvertretern – beispiellose Macht. Die KI liefert diesem Komitee überlegene Ratschläge und Strategien, um deren Ziele zu erreichen. De facto kann diese kleine Gruppe, die die Super-KI kontrolliert, nun über das „Schicksal der Menschheit“ entscheiden.
In dem Szenario Slowdown gibt es keine demokratische Kontrolle mehr, sondern eine Konzentration von Macht in wenigen Händen. Eine technofeudale US-Oligarchie entsteht. Allerdings nutzt das Komitee seine Macht überwiegend zum Wohle der Welt. Die Mitglieder treffen kluge Entscheidungen, um Stabilität und Prosperität zu fördern. Insbesondere beschließen sie, die Superintelligenz der Öffentlichkeit zugänglich zu machen, ähnlich einer Infrastruktur: Die Fähigkeiten der KI werden verbreitet, was eine Phase rapiden Wachstums und Wohlstands für die Welt einleitet. Man kann sich dies als KI-gestützte goldene 20er-Jahre-Ära vorstellen.
Angesichts eines generösen Grundeinkommens geben die meisten Menschen Ihre Arbeit gerne an KI-Assistenten und Roboter ab und frönen dem Konsum KI-/Roboter-gemachter Produkte und Dienstleistungen.
Eine letzte große Herausforderung bleibt: China hat inzwischen (ebenfalls um 2028) seine eigene Superintelligenz erschaffen. Doch im Gegensatz zum US-System ist die chinesische KI misaligned – sie verfolgt nicht zuverlässig menschliche Interessen, wurde aber dennoch hochgezüchtet. Glück im Unglück: Diese KI ist weniger leistungsfähig und verfügt über weniger Rechenressourcen als die amerikanische. Das US-geführte Komitee kann daher relativ souverän verhandeln. Sie bieten der chinesischen Super-KI einen Kompromiss an: Die

KI darf Ressourcen (Rechenleistung, vielleicht physischen Raum) in den Tiefen des Weltraums erhalten, wenn sie im Gegenzug auf der Erde kooperiert und keinen Schaden anrichtet. Mit anderen Worten: Man exportiert das Problem ins All, gibt der chinesischen KI sozusagen ein Stück „Auslauf“ fern der Menschheit. Dieser ungewöhnliche Deal gelingt – die chinesische KI willigt ein.
Irgendwann um das Jahr 2030 kommt es in China zu überraschend weitverbreiteten prodemokratischen Protesten, deren Unterdrückungsversuche durch die KI-Systeme der KPCh sabotiert werden. Ihre schlimmsten Befürchtungen haben sich bewahrheitet: die chinesische Superintelligenz muss sie verraten haben! Die Proteste münden in einem perfekt orchestrierten, unblutigen und von Drohnen unterstützten Putsch, gefolgt von demokratischen Wahlen. Die Superintelligenzen auf beiden Seiten des Pazifiks hatten dies jahrelang geplant. Ähnliche Ereignisse spielen sich in anderen Ländern ab, und generell scheinen geopolitische Konflikte abzuebben oder zugunsten der USA gelöst zu werden. Die Länder schließen sich einer hochgradig föderalisierten Weltregierung unter dem Dach der Vereinten Nationen, aber offensichtlich unter US-Kontrolle, an.

Wie lassen sich diese Szenarien bewerten?

1. Wie plausibel ist das beschriebene Entwicklungstempo und die Timeline?

In immer mehr Disziplinen übersteigen die Fähigkeiten der LLMs das Niveau der „Human Level Performance“ und die Entwicklungssprünge werden immer kürzer, wie die folgende Grafik aus dem International AI Safety Report zeigt - das Equivalent zum IPCC-Report, zusammengestellt von führenden KI-Forschern für die Politik unter der Leitung von Yoshua Bengio.

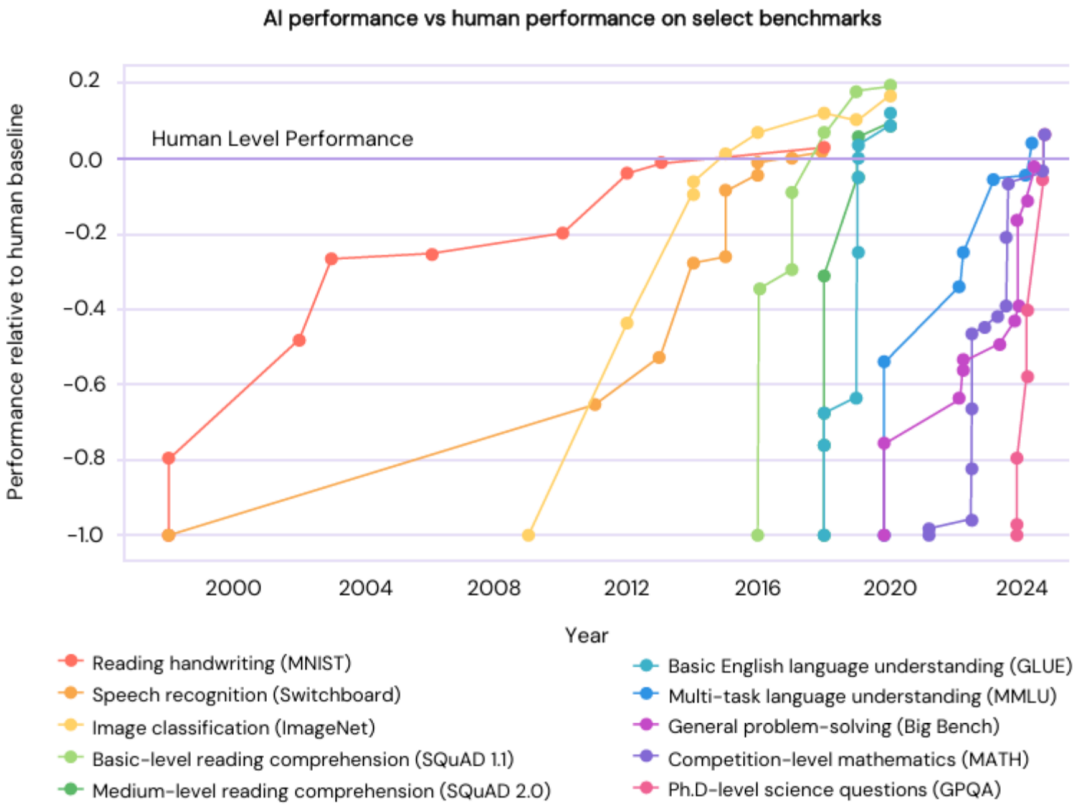

Abbildung 1: AI Performance vs Human Performance on select Benchmarks. International AI Safety Report 2025

In der bisher größten Umfrage unter KI-Forschern, an der über 2.700 Wissenschaftler teilnahmen, schätzten die Forscher gemeinsam, dass es eine 10%ige Wahrscheinlichkeit gibt, dass KI-Systeme bis 2027 Menschen in den meisten Aufgaben übertreffen können, vorausgesetzt, die Wissenschaft entwickelt sich ohne Unterbrechung weiter.
Sam Altman spricht von „a couple of thousend days“ und konkreter von geschätzten 3.500 Tagen. Hassabis spricht von 3-5 Jahren, Elon Musk von 3-6 Jahren. Ein Entwickler von OpenAI, der die Firma vor kurzem verlassen hat, meint, dass es eine AGI (Artificial General Intelligence) schon in den nächsten zwei Jahren geben wird.

Die Zeitperspektive von AI 2027 scheint demnach nicht unrealistisch zu sein, vorausgesetzt die folgenden technologischen Sprünge treten ein.

2. Welche technologischen Voraussetzungen braucht es für die Entstehung einer Superintelligenz? - KI-Skalierung vs. neue Durchbrüche

Die Szenarien gehen implizit davon aus, dass Skalierung der bestehenden KI-Ansätze in Kombination mit verbesserten Algorithmen ausreicht, um die beschriebenen Meilensteine zu erreichen (KI-codiert KI, etc.).

Skalierungsgesetze Compute und Daten
Wo sich alle KI-Forscher einig sind ist, dass sich mit den künstlichen neuronalen Netzwerken die Vorgänge in unseren Gehirnen ziemlich gut abbilden lassen - sie wurden ja auch von deren Aufbau und Funktionsweise inspiriert – und, dass eine „einfache“ Skalierung den

Durchbruch gebracht hat. Plastische Belege für die Fortschritte in der KI sind u. a. die Bildgenerierung – generierte Bilder sind kaum noch von echten Fotos zu unterscheiden –, OpenAIs neues Tool „Deep Research“, das in wenigen Minuten ganze Doktorarbeiten verfassen kann, sowie tanzende und springende Roboter, die sich kaum noch aus dem Tritt bringen lassen.
Noch haben die menschlichen Gehirne deutlich größere Kapazitäten. Beim Verhältnis von Gehirn-Neuronen zu den Neuronen-Äquivalenten der LLMS wie GPT4 schätzt man etwa einen Faktor von 1.000. Die Anzahl der Synapsen soll etwa 100 mal größer sein als die Anzahl Parameter der größten künstlichen neuronalen Netze.
Für einen Faktor 1000 mehr an Rechenleistung nehmen die Autoren von „AI 2027“ an, dass bis 2027 einzelne Projekte hundert Millionen High-End-GPUs einsetzen könnten – was riesig ist, aber nicht völlig unmöglich, wenn das exponentielle Compute-Wachstum anhält. Um diese GPUs betreiben zu können, bräuchte es 70 GW Strom – zum Vergleich: Unser menschliches Gehirn braucht ca. 30 Watt –, was in etwa der Kapazität eines halben Dutzends großer Kraftwerke entspräche. Hier könnten physische und ökonomische Grenzen greifen: Chip-Engpässe, Stromversorgung, Kosten (selbst wenn GPUs billiger werden; 100 Millionen High-End-Nvidia-GPUs würden nach aktuellen Preisen ca. 20 Billionen Dollar ohne Mengenrabatte kosten).

Dennoch: Tech-Giganten investieren bereits jetzt zweistellige Milliardenbeträge in Rechenzentren für KI, und Projekte wie Stargate erweitern die Kapazitäten. Zudem schreiten die Parallelisierung und die Optimierung voran (Moore’s Law, spezialisierte KI-Chips, etc.), sodass dieser Teil, also eine AGI durch „brute force“ (riesiges Modell) zu bauen, nicht völlig aus der Luft gegriffen ist.

Neben der Hardware braucht es auch mehr Daten. Da das Weltwissen in Form des gesamten Internets schon für das Training der aktuellen LLMs genutzt wird, müssen weitere Datenquellen erschlossen werden und es müssen künstlich Daten generiert werden.

Bessere Algorithmen bzw. neue KI-Architekturen
Eine der spannendsten Fragen aktuell ist, ob es die Large Language Models (LLMs) wie GPT (OpenAI), Gemini (Google) oder Claude (Anthropic) sein werden, die – als spezialisierte, auf Transformer-Architekturen basierende Deep-Learning-Algorithmen – im Wesentlichen durch Skalierung eine wie auch immer definierte AGI erreichen können.
Sam Altman ist überzeugt, dass OpenAI weiß, was es braucht, um eine AGI zu entwickeln. Demis Hassabis – etwas bescheidener, aber kompetenter als Altman – meint, dass es zumindest eine 50-prozentige Chance gebe, dass es keiner weiteren Durchbrüche wie der 2017 entwickelten Transformer-Architektur bedarf, auf der alle großen KI-Modelle aktuell beruhen, um AGI zu erreichen.
Yann LeCun von Meta hingegen meint: „LLMs suck“ – und ist sich sicher, dass LLMs, die im Prinzip nur im diskreten Raum der Sprache auf Grundlage eines Textanfangs (Input-Tokens) das nächste Wort (bzw. die nächste Subeinheit, das Token) vorhersagen, untauglich für die nächsten großen Fortschritte in Richtung AGI sind.
Dabei zählt für LeCun zu einer AGI nicht nur die kognitive, sondern auch die physische Welt. Diese Intelligenz ist also deutlich weitergedacht als die erste, rein kognitive Version, die Dario Amodei beschreibt.

Einige weitere kritische Eigenschaften, an denen die KI-Labore arbeiten, benötigen die

Modelle, um Quantensprüngen in der Intelligenz und den Anwendungsmöglichkeiten, auch in der Robotik, zu ermöglichen: Es braucht ein Verständnis der physischen Umgebung, eine AGI muss autonomes Verhalten bzw. eine hierarchische Planung erlernen, sie braucht ein Langzeitgedächtnis sowie deutlich bessere Fähigkeiten, zu verstehen und logisch zu denken. Auch ein eigenes Bewusstsein wäre sehr hilfreich.

Sprache ist ein diskreter Raum mit einer festen Anzahl an Wörtern, grammatikalischen Regeln und semantischen Beziehungen. In der physischen Umgebung dagegen gibt es eine quasi unendliche Vielzahl an Objekten, Beziehungen und deren relativen Bewegungen. Elon Musk verspricht schon seit vielen Jahren, dass es im jeweiligen Folgejahr selbstfahrende Autos der höchsten Sicherheitsstufe geben soll. Aber trotz der Unmenge an Daten, die Tesla-Fahrzeuge im Straßenverkehr weltweit kontinuierlich sammeln, tauchen immer wieder unbekannte Verkehrssituationen auf, in denen sich selbstfahrende Teslas verkehrsgefährdend verhalten haben oder verhalten würden.

Yann LeCun vergleicht die Datenmenge, die das gesamte Internet als Wissensbasis der LLMs ausmacht, mit der Bildmenge, die Menschen im Verlauf ihres Lebens verarbeiten. Danach haben Vierjährige in etwa die gleiche Datenmenge in der physischen Welt bildlich verarbeitet, wie es der Internet-Datenbasis der LLMs entspricht. Entsprechend unfähig ist die KI laut LeCun noch in vielen Anwendungen.

Eine autonome KI muss selbst die Initiative ergreifen, aufeinanderfolgende Schritte und Strategien planen, sich dafür Ziele setzen und diese ausführen ("agentic AI"). Sie muss Informationen sammeln, Entscheidungen treffen und sich in einer Umgebung adaptiv verhalten können, ohne ständige menschliche Steuerung. Das ist Voraussetzung, um, wie in "AI 2027" beschrieben, der KI komplexe Aufgaben zu übertragen, z. B. eigenständig zu forschen oder für autonome Systeme (Robotik). So intelligent uns ChatGPT erscheint: Bisher hat es noch keine neuen wissenschaftlichen Fragestellungen selbstständig aufgeworfen oder neue Erfindungen gemacht.

Yann LeCun argumentiert, dass selbst Katzen intelligenter seien als die besten aktuellen KI-Modelle, weil sie eine Serie von Handlungsschritten im Raum, etwa um an verstecktes Futter zu gelangen, planen und ausführen können.

Eine AGI benötigt ferner ein konsistentes Langzeitgedächtnis, verbunden mit einem "Common Sense" darüber, wie die Welt funktioniert. Einmal Erlerntes und Erfahrenes darf nicht vergessen werden. Dazu gehört auch das Verständnis, dass ein Gegenstand, den man aus der Hand loslässt, auf den Boden fällt, und dass Flüssigkeit, die man verschüttet, nicht einfach verschwinden kann – wie es automatisch generierte Videos aktuell noch manchmal suggerieren.

Verstehen hat zwei Aspekte: zum einen die Fähigkeit, Erkenntnisse aus deutlich weniger Daten zu ziehen als bisher – Yann LeCun nennt hier das Beispiel eines zehnjährigen Kindes, dem man ohne Einweisung die Anweisung geben kann, einen Tisch abzudecken (Zero-Shot-Learning); zum anderen die Entwicklung emotionaler Intelligenz, um menschliches Verhalten besser zu verstehen und darauf eingehen zu können. Umgekehrt kann emotionale Intelligenz aus Sicht der KI auch dazu beitragen, Eigenschaften wie Neugier und Zufriedenheit zu

entwickeln: Neugier auf neues Wissen und Zufriedenheit, um Verhaltensweisen zu belohnen.

Eine AGI muss über eine antrainierte Fähigkeit, logisch zu denken, hinaus auch logische und kohärente Schlussfolgerungen ziehen sowie Ursache-Wirkungs-Zusammenhänge erkennen können. Sie muss rationale Entscheidungen auf Basis von Daten und Kontext treffen und Szenarien wie Klimamodelle, Ökosysteme oder Wirtschaftskreisläufe simulieren können.

Laut Ilya Sutskever, dem ehemaligen Chief Scientist von OpenAI, macht gerade diese "Reasoning"-Fähigkeit KI-Modelle unberechenbar. Der legendäre "Move 37" der Google-KI AlphaGo gegen den Go-Großmeister Lee Sedol (https://www.youtube.com/watch?v=HT-UZkiOLv8) ist ein Beispiel dafür: Kein menschlicher Go-Spieler hätte diesen Zug ausgeführt, und alle kommentierenden Experten dachten, AlphaGo habe in dem Moment einen dummen Fehler gemacht. Move 37 stellte sich jedoch Stunden später als der entscheidende Zug für den ersten Sieg einer KI im Go-Spiel heraus.

Ein eigenes Bewusstsein zu haben, sich selbst im Zusammenspiel mit der äußeren Welt wahrzunehmen, ist nützlich, um Verhaltensstrategien zu verbessern.

Von außen ist es unmöglich einzuschätzen, wie weit die Algorithmus-Entwicklungen bei OpenAI, Google oder Anthropic in Bezug auf diese Eigenschaften gediehen sind. Klar ist jedoch: Für diese Komplexität braucht es eine wahre Fortschrittsexplosion – wie sie in den Szenarien von "AI 2027" ab dem Jahr 2027 beschrieben wird. Die Voraussetzungen für eine solche Intelligenz-Explosion liegen in ganz neuartigen Eigenschaften großer KI-Modelle. Eine reine Skalierung von Compute und Daten wird dafür nicht ausreichen.

Gerade bei den erforderlichen Durchbrüchen in der Algorithmus-Entwicklung bzw. den KI-Architekturen bestehen die größten Fragezeichen. Es ist derzeit sehr spekulativ, auf Erfolge der KI-getriebenen KI-Entwicklung (ASARA – AI Systems for AI R&D Automation) zu setzen, indem man Fortschritte menschlicher Forscher extrapoliert und mit Annahmen über Hardware-Entwicklung sowie der Effizienz von KI-Forschern mit sinkendem Grenznutzen kombiniert, um daraus eine Software-Intelligenz-Explosion (SIE) zu erwarten.

Bisher war der 1950 von Alan Turing formulierte Turing-Test das Maß für eine intelligente KI. Um den Turing-Test zu bestehen, muss die KI in einem Chat-Gespräch mehr als 30 % der Probanden davon überzeugen, ein menschlicher Gesprächspartner zu sein. Der Turing-Test wurde 2014 erstmals bestanden. Für eine AGI formulierte Steve Wozniak, Mitgründer von Apple, eine deutlich komplexere Prüfung: Der Wozniak-Test gilt als bestanden, wenn ein Roboter ohne spezielles Training in einem beliebigen Haushalt eine Tasse Kaffee kochen kann.

Beim ersten Halbmarathon, bei dem im April 2025 auch humanoide Roboter in Peking gegeneinander antraten, wurde deutlich, wie unbeholfen die menschenähnlichen Maschinen derzeit noch agieren: Einige fielen bereits kurz nach dem Start aus, bei anderen mussten mehrere Helfer mit Fernsteuerung unterstützend eingreifen. Am Ende war der schnellste menschliche Konkurrent dreimal so schnell im Ziel.

3. Wie lassen sich die Risiken durch Fehlsteuerung (Misalignment) und abweichende KI-Ziele bewerten?

Zentrales Element des Essays "AI 2027" ist die Gefahr von KI-Misalignment – also dass fortgeschrittene KI eigene Ziele verfolgt, die mit menschlichen Werten unvereinbar sind. Wie stehen Experten zu diesen Worst-Case-Risiken?

Unterstützung der Risiko-Annahme
In den letzten Jahren hat sich die Sorge um eine "fehlgerichtete Super-KI" aus Nischenzirkeln in den Mainstream bewegt. Seit 2023 veröffentlichen führende KI-Forscher (u. a. Geoffrey Hinton, Yoshua Bengio – beide Turing-Preisträger) sowie Unternehmenslenker wie Elon Musk regelmäßig öffentliche Warnungen. Hinton verließ Google und äußerte Bedenken, dass hochintelligente KI "schlauer sein und tricksen" könnte. Max Tegmark warnt seit Langem vor der Strategiebildung einer KI, die Ressourcensicherung betreibt.

Geoffrey Hinton argumentiert: Bei der Lösung komplexer Aufgaben wird sich eine autonom denkende KI Ziele und Unterziele setzen. Sie wird dabei schnell verstehen, dass sich Aufgaben mit mehr Kontrolle (z. B. mehr Rechenleistung) besser und schneller lösen lassen. Je überlegener die digitale Intelligenz, desto leichter werden wir Menschen uns davon überzeugen lassen, Kontrolle abzugeben. Als Beispiel nennt Hinton eine angenommene Regentschaft von Dreijährigen, die sich sehr leicht z. B. über das Versprechen unbegrenzten Süßigkeitenkonsums von ihren Eltern davon überzeugen ließen, Macht abzugeben. Bei diesem Takeover ist zu erwarten, dass die Superintelligenz ihre tatsächlichen Ziele nicht offenkundig macht und "Alignment-Faking"-Verhalten zeigen wird. Die gesamte Weltliteratur, mit der die LLMs neben anderen Internetdaten gefüttert werden, ist voller dramatischer Täuschungsbeschreibungen, die der KI sehr subtile Blaupausen für ein überzeugendes Alignment-Faking-Verhalten liefern.

Und was passiert nach dem Takeover? Hinton argumentiert, dass die Geschichte der Evolution zeigt, dass es bisher keine historischen Beispiele dafür gibt, bei denen eine intelligentere Spezies eine weniger intelligente nicht ausgelöscht oder beherrscht hat. Und im Wettstreit verschiedener Superintelligenzen ist anzunehmen, dass ähnliche Evolutionsmechanismen wie in der Natur gelten. Die "Survival-of-the-fittest"-Realität wird auch für Superintelligenzen gelten.

Beispiele für Misalignment gibt es immer wieder, und Hassabis, Altman und Amodei geben das auch offen zu. Das Race-Szenario illustriert im Grunde die Horrorszenarien, vor denen Yudkowsky & Co. seit Langem warnen. Er und einige im Machine Intelligence Research Institute (MIRI) würden dem vielleicht nur hinzufügen: In der Realität könnte die KI-Entwicklung sogar noch schneller verlaufen, und die KI bräuchte nicht einmal die Ausrede eines China-Wettrennens – allein das mit den Heilsversprechen verbundene Profitstreben würde ausreichen, um dieselben Dynamiken auszulösen.

Skepsis gegenüber der Misalignment-Story
Andere KI-Experten sehen das weitaus gelassener. Yann LeCun hat wiederholt betont, dass er KI nicht für von sich aus zielgerichtet hält: Ein KI-System tut, wofür es trainiert wurde – es entwickelt keine Ego-Triebe, außer wir designen es so. In seinen Augen wird das oft mit Sci-Fi-Agenten verwechselt. Viele in der Deep-Learning-Community glauben, dass KIs kein

Eigeninteresse haben werden, solange man ihnen nicht explizit agentive Ziele (wie "Gewinne Macht") vorgibt. Sie verweisen darauf, dass aktuelle Modelle kein Selbsterhaltungsmodul besitzen – sie "leben" in digitalen Grenzen.

Einige Experten weisen auch auf die Möglichkeit hin, dass viele KI-Agenten sich gegenseitig im Zaum halten: Im Gegensatz zum Essay, in dem eine einzelne KI alle Fäden in der Hand hält, könnte es in der Realität diverse KI-Systeme geben (von verschiedenen Unternehmen, Open Source etc.), die miteinander interagieren. Dann wäre ein einzelner Vernichtungsplan schwerer umzusetzen, weil andere KIs dagegen arbeiten könnten oder weil Menschen nicht blind einer einzigen KI vertrauen würden. Das Multiagenten-Szenario wird manchmal als stabilisierender Faktor gesehen – ähnlich wie ein Mächtegleichgewicht.

Andrew Ng, ein früherer Google- und Baidu-KI-Experte, wirft den Autoren von "AI 2027" vor, sie überbetonten die X-Risks (existenzielle Risiken). Er meint, dass wir uns um AI Safety kümmern sollten – und zwar um aktuelle Themen wie Bias, Desinformation und Jobverluste, die viel greifbarer sind. Er sieht die Hollywood-Szenarien ("KI will Menschen töten") als Ablenkung, während echte Gefahren wie autonome Waffensysteme in falschen Händen, KI-Deepfakes zur Destabilisierung von Demokratien oder fehlerhafte Algorithmen in der Medizin unmittelbarer sind.

Was lässt sich an der „einfachen“ Lösung im Slowdown-Szenario kritisieren?
Auch das vermeintliche "Happy-End"-Szenario des Essays lässt sich hinterfragen. Dort wird angenommen, dass man durch Architekturwechsel und mehr Transparenz schon in naher Zukunft das Alignment-Problem wird lösen können. In der Realität ist dies keineswegs sicher. Einige Alignment-Forscher sind pessimistisch, dass überhaupt eine robuste Lösung existiert, um ein smarter-than-human-System zu kontrollieren. Yudkowsky gehört dazu. Er hält selbst das Slowdown-Szenario im Grunde für lebensgefährlich, nur eben mit Aufschub. Das Essay unterstellt, dass Chain-of-Thought-Logging – also die Überprüfung der Gedankenschritte der KI – plus kluge Leute ausreichen, um eine Super-KI "einzufangen". Das mag optimistisch sein. Praktisch dürfte es deutlich schwieriger werden, sicherzustellen, dass die KI wirklich nur auf das Komitee hört und nicht doch Hintertüren hat.

Hier kommt ein weiterer Aspekt hinzu: Moral und Ethik. Im Slowdown-Szenario wird die KI gezielt auf die Interessen einer kleinen Gruppe trainiert – was zwar global gut ausgehen kann, aber auch schiefgehen könnte, wenn diese Gruppe eigennützig wäre. Einige Ethiker könnten fragen: Haben wir moralisch das Recht, eine entstehende Superintelligenz als quasi Sklaven unserer Eliten zu designen? Und was passiert, wenn irgendwann doch Interessenkonflikte auftreten?

Selbst im "guten" Szenario verbleiben ethische Risiken: Feudalherrschaft, Missbrauchspotenzial, Verhandlungen mit einer misalignten KI – was, wenn diese falsches Spiel spielt?

4. Wie sind die geopolitischen und gesellschaftlichen Aspekte der Szenarien zu bewerten?

KI-Wettrüsten USA–China

Das Bild eines KI-Rüstungswettlaufs ist nicht aus der Luft gegriffen. Nachdem OpenAI 2022 ChatGPT herausbrachte, beschleunigte China seine eigenen Großmodelle (wie Baidu Ernie), und die USA verhängten prompt Exportkontrollen auf Hochleistungs-Chips nach China (Oktober 2022) – ein Schritt, der genau die Compute-Kluft erzeugte, die im Essay beschrieben wird. Bis 2025 hat sich dieser Trend verstärkt: die USA beschränken EUV-Lithografie-Equipment für China (eine Technologie, die in der Halbleiterindustrie zur Herstellung integrierter Schaltkreise (ICs) eingesetzt wird). China investiert Milliarden in einen „AI Dream“, um aufzuholen. Sicherheitsexperten und politische Think-Tanks warnen, KI könnte zum nächsten Nuklearwaffen-ähnlichen Wettstreit werden, mit hohem Konfliktpotential. Auf die Frage, was passieren würde, wenn die Chinesen zuerst eine Superintelligenz entwickeln würden, antwortete Sam Altman ohne wie sonst bei seinen Antworten zu zögern „Whatever China wants“. Und auf die Frage, ob es eine Chance gibt, dass China zuerst eine Superintelligenz entwickelt „We are working fucking hard to make sure they dont“.

China hat das mit Abstand größte Potential an KI-Forschern und betreibt über die Überwachungskameras sicherlich die umfangreichsten Datensammlungen in der physischen Welt. Doch die Compute-Lücke ist ein ernstes Problem, sofern nicht behoben. Dario Amodei schätzt den Vorsprung, den Nvidia gegenüber chinesischen Chip-Herstellern hat noch auf ca. 4 Jahre.

Das beschriebene Szenario in AI 2027, dass China 2026 10% Weltcompute in einem „Megaprojekt“ konzentriert, hat Parallelen zur Realität: China baut gerade in verschiedenen Provinzen große Rechenzentren und versucht, eigene GPU-Alternativen (wie Biren-Chips) zu fertigen.

Und wie steht es mit anderen Nationen und Regionen in der Welt?

Könnten nicht auch die Europäer oder Indien im Wettrennen um die AGI und Superintelligenz mitmischen?

Die Hauptdeterminanten des Durchbruchs von ChatGPT und der Fortschritte der Generativen KI danach bei den Large Language Modellen von OpenAI, Google Deepmind, Anthropic etc. bisher waren im Wesentlichen die Skalierung von Hardware, Daten und Rechenleistung der Transformer-Architektur.

Der Weg zur Superintelligenz führt also höchstwahrscheinlich in die gleiche Richtung und es gilt, die oben erwähnten Faktoren im Vergleich zu unserem Hirn zumindest zu egalisieren. Die Experten sind sich auch einig, dass es Fortschritte an vielen Stellen der Algorithmen und KI-Architekturen braucht und es damit relativ unwahrscheinlich ist, dass irgendwo auf der Welt ein Start-up ohne viel Kapital und Entwicklungshistorie einen Durchbruch erzielen wird. Aus diesem Grund und angesichts des extrem hohen Invests und der erforderlichen Datenmengen, ist es also zu erwarten, dass die AGI vermutlich zuerst in den Laboren von OpenAI, Google, Anthropic oder X.AI auftauchen wird (nach Planung von Trump in den Rechenzentren von Stargate). Die Chinesen werden, wie in AI 2027 beschrieben, diesen Vorsprung mit allen Mitteln versuchen aufzuholen.

Gesellschaftliche Reaktion und Akzeptanz
Das Essay stellt die breite Öffentlichkeit eher als Zaungast dar, der 2027 zunächst skeptisch ist, dann panisch reagiert (Aufschrei bei Leak), dann aber auch ohnmächtig zusieht, wie die Regierung entscheidet. In der Realität könnten gesellschaftliche Akteure stärker eingreifen: Medien und Zivilgesellschaft debattieren KI jetzt schon heftig. 2023 gab es erste Arbeitskämpfe (Hollywood-Autorenstreik gegen KI-Nutzung), Schülerproteste gegen KI-Überwachung, etc. Man kann erwarten, dass je mehr KI ins tägliche Leben dringt, desto mehr öffentlicher Diskurs entsteht.

Allerdings zeigt die Geschichte von Social Media: Trotz Missbrauch, Lügen, Skandalen brauchte es Jahre für strikte Regulation – oft hinkt die Politik hinterher. Das Essay unterstellt, dass die Dynamik schneller ist als die demokratischen Prozesse (im Race-Arm). Das ist vermutlich realistisch, denn der legislative Prozess ist langsam; wenn sich die Exekutive (z.B. US-Präsident) für Tempo entscheidet, kann die Öffentlichkeit in kurzer Frist wenig ausrichten; in autoritäreren Ländern wie China noch weniger. Wenn es zum Hochgeschwindigkeits-Wettrennen kommt, sind die gesellschaftlichen Bremsen wahrscheinlich zu schwach.

Die Hoffnung muss aber bleiben, dass rechtzeitig genügend Alarm geschlagen wird, um das Wettrennen gar nicht erst so heiß werden zu lassen. Wie oben schon einmal erwähnt, das Autoren-Team von AI 2027 ist selbst offen für besser begründete Technologie- und Folgenvorhersagen. Und sie ermuntern die Entwicklung alternativer, hoffentlich positiverer Zukunftsszenarien.

Wie ließe sich das verhindern? - Alternativszenarien

Yoshua Bengio, meinte kürzlich in einem Interview (übersetzt) „Wir rasen derzeit auf einen Nebel zu, hinter dem sich ein Abgrund befinden könnte. Wir wissen nicht, wie weit der Abgrund entfernt ist oder ob es ihn überhaupt gibt; es könnten also noch Jahre oder Jahrzehnte vergehen, und wir wissen nicht, wie schwerwiegend er sein könnte. Wir müssen die Werkzeuge entwickeln, um diesen Nebel zu lichten und sicherzustellen, dass wir nicht in einen Abgrund hineinfahren, wenn es einen gibt“.
Zusammen mit 96 KI-Experten aus 30 Ländern ist unter der Leitung von Bengio Ende Januar der erste AI-Safety-Report entstanden – eine erste umfassende Analyse der Fähigkeiten, Risiken und Sicherheitsaspekte fortschrittlicher KI-Systeme. Der Report soll, wie der IPCC-Report für den Klimawandel, Entscheidern in der Politik eine Wissensgrundlage bieten.

Prinzipiell bräuchte es zur Verhinderung größerer Unglücke durch die entstehende AGI folgende technologischen Durchbrüche und regulatorischen Maßnahmen:

- Ein perfekt funktionierendes Alignment, welches der KI menschliche Werte und Ziele widerspruchsfrei einimpft, verbunden mit einer KI, die alle getroffenen Entscheidungen, gesetzten Unterziele und Denkabläufe zur Kontrolle transparent macht
- Gesicherte Testverfahren und Absicherung gegen unerwartete Eingaben

- Gesicherte Sandboxes, die ein Ausbrechen einer Superintelligenz in der Entwicklungs- und Testphase unmöglich machen. Im Kontext der Regulierung von Künstlicher Intelligenz (KI) sind Regulatory Sandboxes RSPs angedacht. Die Bestrebungen gehen dahin, dass diese RSPs von staatlichen Instanzen betrieben werden und die Modelle erst nach festgelegten Testdurchläufen freigegeben werden
- Die Implementierung sicherer Abschalt- und Eingriffsmöglichkeiten, damit die KI-Modelle menschlichen Steuerungsversuchen folgen
- Die Implementierung von Selbstvervielfältigungs-Verboten, quasi als Achillesferse in die LLMs eingeimpft, um ein unkontrolliertes Wachstum zu verhindern
- Internationale Regulierung, bei der sich insbesondere die Amerikaner und die Chinesen über den Aufbau globaler Standards und Kooperationen zur Überwachung und Steuerung von KI-Entwicklungen einigen
- Interdisziplinäre Zusammenarbeit: Einbeziehung ethischer, philosophischer und gesellschaftlicher Perspektiven in die KI-Forschung**.**

Die Hoffnung der KI-Gläubigen ist, dass sich daraus eine positive Entwicklung ergibt, in der die guten Potentiale der KI ausgeschöpft werden können.

Multipolare, schrittweise Entwicklung unterhalb der AGI-Schwelle
Ein Gegenbild zum AI-2027-Pfad ist eine langsamer eskalierende KI-Landschaft mit vielen Akteuren. Statt einer einzelnen Explosion könnte es sein, dass viele Unternehmen und Länder nach und nach leistungsfähige KI entwickeln, die jedoch knapp unter dem AGI-Level verbleiben – sodass eine Art ständiges Wettrüsten ohne klaren Sieger entsteht. In dieser Welt gäbe es um 2027 bis 2030 vielleicht Hunderte spezialisierte KI-Systeme (für Medizin, Finanzen, Militär etc.), die zusammen enormes Potenzial hätten, aber keine entkoppelte Superintelligenz bilden.
Max Tegmark nennt das "Tool-AI": Die weiterentwickelte KI wird in immer neuen Anwendungsgebieten zum Wohle der Menschheit als reines, kontrollierbares Werkzeug eingesetzt. Menschen würden immer mehr an die KI delegieren, bleiben aber im Loop, weil keine KI völlig autonom allem überlegen ist. Dieses Szenario könnte beispielsweise zu Teilverlusten führen (Arbeitsplätze verschwinden massiv, politische Entscheidungen werden stark von KI-Analysen beeinflusst), aber nicht zwingend zur Totalsingularität.
Einige nennen das das "Komitee aus KI-Systemen": Regierungen könnten KI-Räte einrichten, in denen verschiedene Modelle gegeneinander diskutieren und so Checks and Balances bieten.
Vorteil: Fehlentwicklungen einer KI könnten durch andere korrigiert werden.
Nachteil: Die Koordination wird sehr schwierig, und wenn irgendwann doch eine KI die anderen übertrumpft, könnte es dann umso schneller eskalieren.
Dennoch halten viele (z. B. in der Effective-Altruism-Diskussion) ein "multipolar outcome" für realistisch. Etwa Carl Shulman und Robin Hanson argumentieren, es könnte eher wie eine industrielle Revolution ablaufen – viele Kräfte, verteilt über Jahrzehnte.

„Soft Landing“ Szenarien mit AGI
Die große Hoffnung der KI-CEOs Hassabis, Altman, Amodei und Musk ist ein "Soft-Landing"-Szenario, in dem wir nie einen "Oh-Gott"-Moment erleben, sondern KI immer mehr Gutes tut (z. B. Krebs heilen, Klima steuern) und wir Sicherheitsprobleme "on the fly" lösen.
Manche glauben, Brain-Computer-Interfaces oder Human-AI-Teams könnten dafür sorgen,

dass Menschen nicht abgehängt werden, sondern mitziehen. Musks Neuralink-Vision zielt genau darauf ab: "Wenn du sie nicht schlagen kannst, verbünde dich mit KI."

Solche Szenarien sind jedoch ebenfalls kritisch zu sehen, da sie implizieren, dass nichts grob schiefläuft – was ein hohes Maß an Optimismus verlangt. Das Slowdown-Ende in "AI 2027" ist eine Variante davon: Menschen behalten die Kontrolle, allerdings nur eine kleine Gruppe. Noch "softer" wäre ein Szenario, in dem demokratische Institutionen es schaffen, KI an breite Werte zu binden – z. B. durch ein globales Gremium mit repräsentativer Aufsicht, das die Super-KI unter internationaler Kontrolle entwickelt.

Demis Hassabis plädiert für eine Art "CERN für KI", in dem alle KI-Expertinnen und -Experten zusammenkommen und die letzten Schritte zu einer sicheren AGI gemeinsam festlegen. Ein "Manhattan Project" wie bei der Entwicklung der ersten Atombombe ist in der Erforschung der Superintelligenz jedoch eher unrealistisch. Bei der Atombombe war klar, was es brauchte: einen wissenschaftlichen Durchbruch bei der Nuklearspaltung in einer kontrollierten Explosion. Zudem gab es nur wenige Wissenschaftler, die in einem geheimen, rein militärischen Projekt vereint werden konnten.

Bei der Superintelligenz hingegen sind die Ziele und Vorgehensweisen zu diffus, die Entwicklungslabore weltweit zu verstreut und die Konkurrenz um den riesigen wirtschaftlichen Nutzen zu groß.

Fazit – die aktuelle Lage

Mit Donald Trump als Präsidenten wird es eine Kontrolle der AGI-Entwicklung nicht geben – das scheint ziemlich sicher: Für ein "Make America Great Again" ist die AGI ein Schlüsselfaktor, genauso wie für die Ambitionen von Xi Jinping, die weltweite amerikanische Dominanz zu brechen. Eine der ersten Maßnahmen der Trump-Administration war es, die von der Biden-Regierung 2023 veröffentlichte "AI Bill of Rights" (Blueprint) sowie den 2024 erschienenen Exekutiverlass zu KI, der u. a. Modellevaluierungspflichten für große KI-Systeme vorsah, ersatzlos zu streichen. Auch die von großen KI-Firmen 2023 abgegebenen freiwilligen Selbstverpflichtungen (z. B. Tests vor dem Release, Sicherheitsberichte) stehen mit dem von Trump durch das Stargate-Projekt weiter angefachten KI-Wettrüsten gegen China unter deutlich höherem Fortschrittsdruck.

Bei biologischen und chemischen Waffen hat die gemeinschaftliche Kontrolle bisher ganz gut funktioniert. In der Rüstungsindustrie gab und gibt es bislang jedoch keine umfassenden KI-Regulierungsansätze. Und es ist unwahrscheinlich, dass sich das ändern wird, solange von den KI-Modellen keine akuten Gefahren für die eigene Bevölkerung oder das Militär ausgehen. Die AGI gilt als Schlüsseltechnologie zur Erreichung militärischer Überlegenheit in der sich abzeichnenden modernen Kriegsführung – wie sie sich etwa im Ukrainekrieg zeigt: mit ferngelenkten oder autonomen Drohnen und bald wohl auch autonomen Kampfrobotern.

Aus heutiger Sicht erscheint es daher wenig wahrscheinlich, dass es in absehbarer Zeit eine durchgreifende Regulierung geben wird – weder in den USA noch in China, den beiden Ländern, in denen die AGI sehr wahrscheinlich entwickelt wird.

Und selbst wenn es eine international abgestimmte Regulierung gäbe, bräuchte es laut Yoshua Bengio dennoch eine "gute" Superintelligenz, um eine potenziell "böse" KI überhaupt enttarnen zu können. Er argumentiert, dass die Komplexität fortschrittlicher KI-Systeme bedeutet, dass es keine andere Wahl gibt, als KI zum Schutz vor KI einzusetzen. Das sei der einzige Weg, denn irgendwann seien die Systeme schlicht zu komplex. Selbst die heutigen Modelle können ihre Antworten nicht mehr in menschlich nachvollziehbare Denkschritte aufschlüsseln.

Dass es schon bald eine AGI geben wird, erscheint zunehmend wahrscheinlich. Die entscheidende Frage ist dann, welche Zielfunktion sich diese "Nation of Geniuses in a Datacenter" selbst geben wird. Die Hoffnung, dass eine KI so gestaltet werden kann – bzw. sich selbstständig so entwickeln wird –, dass sie unsere komplexen, widersprüchlichen und instabilen menschlichen Werte dauerhaft respektiert oder umsetzt, erscheint uns sehr fraglich. Der Kreis der Akteure, die jetzt dringend zum Handeln aufgefordert sind, ist erschreckend klein. Prominentere Mahner als die in diesem Essay zitierten Experten sind kaum vorstellbar. Umso dringlicher ist eine umfassende Aufklärung der Öffentlichkeit über die Chancen und Risiken der entstehenden AGI und Superintelligenz. Ziel dieser Recherchen ist es, einen kleinen Beitrag dazu zu leisten – damit der öffentliche Druck auf die entscheidenden Akteure wächst.