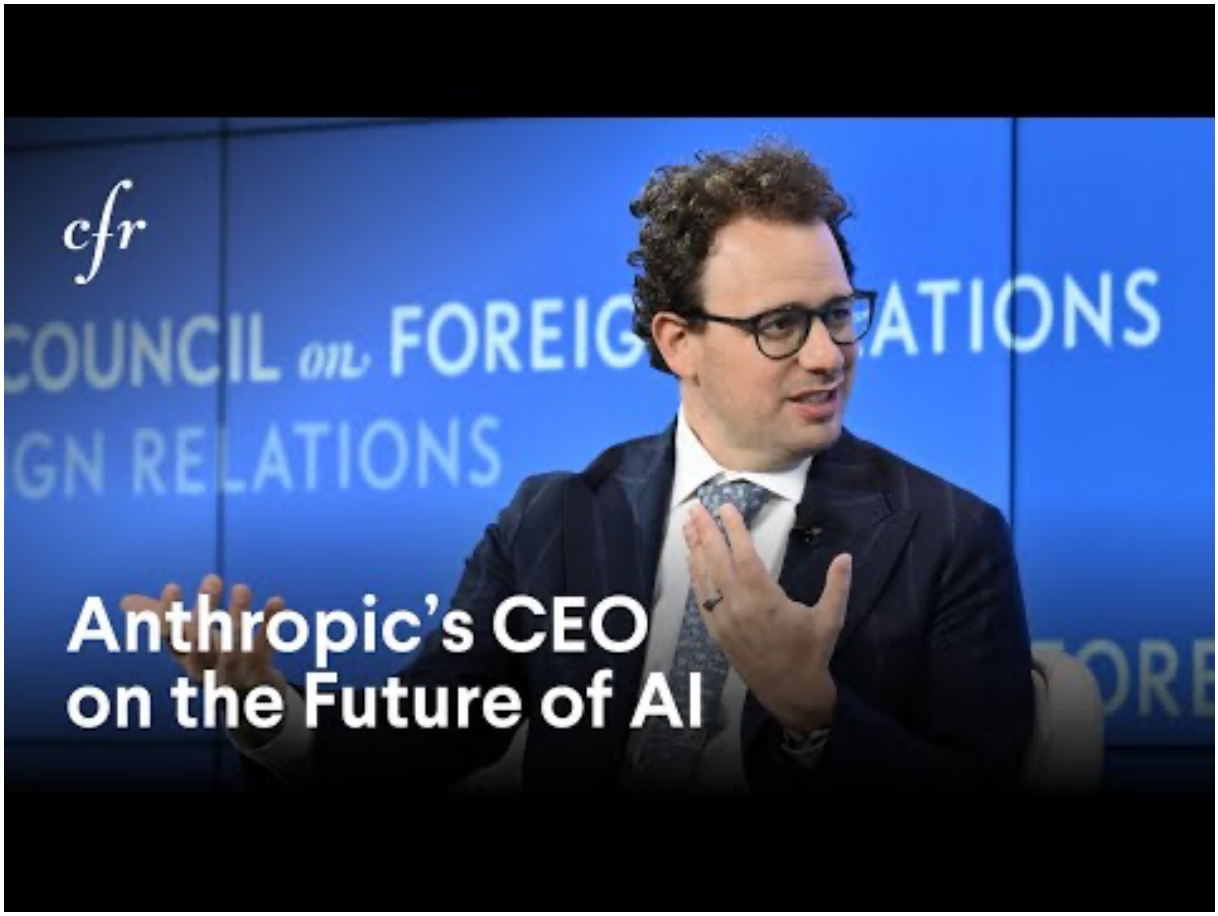
Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Video vom 12.03.2025
https://www.youtube.com/watch?v=esCSpbDPJik



Zusammenfassung
In diesem aufschlussreichen Gespräch interviewt Mike Froman Dario Amodei, den **CEO und Mitbegründer von Anthropic, über den Stand der künstlichen Intelligenz (KI), ihre Auswirkungen auf die Gesellschaft und die Vision hinter Anthropic als gemeinnütziges Unternehmen**. Amodei diskutiert seinen Abschied von OpenAI und führt ihn zum Teil auf die Anerkennung der **Skalierungshypothese zurück, die besagt, dass KI-Modelle mit zunehmender Rechenleistung und Trainingsdaten eine grössere Intelligenz bei verschiedenen Aufgaben erreichen.**
Er betont die Notwendigkeit einer verantwortungsvollen Entwicklung und Sicherheit von KI-Technologien und skizziert, wie Anthropic sich für Werte wie mechanistische Interpretierbarkeit und konstitutionelle KI einsetzt. Anhand von Beispielen wie der Verzögerung der Produktfreigabe für Sicherheits-bewertungen und der Umsetzung einer verantwortungsvollen Skalierungs-richtlinie unterstreicht die Diskussion das Engagement von Anthropic, sicherzustellen, dass KI dem öffentlichen Interesse und nicht nur kommerziellen Gewinnen dient.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Amodei untersucht ausserdem die Risiken von KI, einschliesslich Bedenken hinsichtlich der nationalen Sicherheit und möglicher Arbeitsplatzverdrängung, und erörtert auch die Aufwärtspotenziale, die KI bietet, insbesondere in Bereichen wie dem Gesundheitswesen.
Er prognostiziert eine Zukunft, in der KI-Branchen revolutionieren, die Produktivität steigern und sogar bei der Lösung komplexer Krankheiten helfen könnte.
Er wirft jedoch existenzielle Fragen über die Rolle des Menschen in einer Welt auf, in der KI ihn in vielen Aufgaben übertreffen könnte.
Der Dialog befasst sich mit der **nationalen Sicherheit, Exportkontrollen und der Perspektive von KI in den Sozialwissenschaften**.
Letztendlich plädiert Amodei für einen sorgfältigen Ansatz beim Management der KI-Entwicklung und betont **die Bedeutung menschlicher Verbindungen und Ambitionen** inmitten steigender technologischer Fähigkeiten.

Höhepunkte

- 🤖 Die Mission von Anthropic: Engagement für einen missionsorientierten Ansatz in der KI, **der sich von traditionellen gewinnorientierten Modellen unterscheidet.**
- 🔍 **Mechanistische Interpretierbarkeit**: Erforschung des Innenlebens von KI, um Transparenz und Sicherheit zum Nutzen der Gesellschaft zu gewährleisten.
- x **Verantwortungsvolle-Skalierung**: Ein Framework, das Sicherheitsprotokolle priorisiert, da KI-Modelle immer fortschrittlicher werden.
- 🌐 KI als zweischneidiges Schwert: Abwägung der wirtschaftlichen Vorteile von KI gegen das **Potenzial für eine grossflächige Verdrängung von Arbeitsplätzen.**
- 🏥 **Transformative Gesundheitslösungen**: Spekulationen über das Potenzial von KI, das Gesundheitswesen zu revolutionieren und komplexe Krankheiten zu bekämpfen.
- ⚔️ **Auswirkungen auf die nationale Sicherheit**: Verstehen, wie KI-Technologie die globale Machtdynamik und Sicherheitsbedrohungen beeinflusst.
- 👥 **Human Experience**: Reflexion darüber, was es bedeutet, in einer Ära fortschrittlicher KI ein Mensch zu sein, und wie wichtig es ist, sinnvolle Beziehungen zu pflegen.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Wichtige Erkenntnisse

- 🔍 **Einblicke in die Skalierungshypothese**: Amodei hebt die **"Skalierungsgesetze"** hervor, die darauf hindeuten, dass Verbesserungen in der Leistung von **KI-Modellen direkt mit mehr Rechenleistung und Daten korrelieren**. Dies deutet auf einen Paradigmenwechsel bei den KI-Fähigkeiten hin, der zu einer Neubewertung der nationalen Sicherheit, ethischer Erwägungen und gesellschaftlicher Auswirkungen führt.

- 📖 **Mechanistische Interpretierbarkeit**: Das Engagement von Anthropic, das "Warum" hinter KI-Entscheidungen zu verstehen, unterstreicht die Bedeutung transparenter KI-Systeme. Durch die Investition in diese Forschung setzt sich Anthropic für eine sachkundige Gesellschaft ein, die die Vorteile der KI besser nutzen und gleichzeitig unvorhergesehene Folgen abmildern kann.

- ⏳ **Verzögerung der Produktfreigabe aus Sicherheitsgründen**: Durch die Entscheidung für eine verzögerte Markteinführung ihres KI-Modells zeigte Anthropic eine Kultur**, in der der Sicherheit Vorrang vor unmittelbaren kommerziellen Gewinnen eingeräumt wird.** Dieser Ansatz könnte als kritischer Präzedenzfall in der Technologiebranche dienen, in der eine schnelle Bereitstellung oft ethische Überlegungen überschattet.

- ⚖️ **Constitutional AI Framework**: Dieser innovative Ansatz für das Training von Modellen dreht sich um Leitprinzipien und nicht nur um datengetriebene Methoden. Es positioniert Anthropic in einem einzigartigen Raum, in dem Verantwortlichkeit etabliert werden kann, und schliesst einen Vertrag mit der Gesellschaft darüber, was die Technologie erreichen soll.

- 👥 **Risiken von KI bei der Verdrängung von Arbeitsplätzen**: Amodei erkennt die Komplexität des Arbeitsmarktes an, da KI in der Lage wird, menschliche Aufgaben zu erledigen. Die Folgen der weit verbreiteten Vertreibung erfordern Weitsicht und eine durchdachte Politikgestaltung, um soziale Stabilität und Werte zu erhalten.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

- **KI im Gesundheitswesen**: Die potenziellen Auswirkungen von KI in der medizinischen Wissenschaft stellen eine hoffnungsvolle Vision für die Lösung chronischer Gesundheitsprobleme dar. Die Nutzung von KI kann zu beschleunigten Behandlungen und Verbesserungen der globalen Gesundheitsergebnisse führen, was die transformativen Möglichkeiten dieser Technologie zeigt.

- **Nationale Sicherheit und Exportkontrollen**: Das Gespräch dreht sich um die Auswirkungen neuer KI-Technologien auf die nationale Sicherheit und plädiert für strategische Exportkontrollen, um zu verhindern, dass sich gegnerische Länder einen Vorteil bei der KI-Entwicklung verschaffen. Dies unterstreicht die Verflechtung von technologischer und militärischer Dominanz in der heutigen globalen Landschaft.

Schlussfolgerung

Der Dialog zwischen Mike Froman und Dario Amodei verkörpert die zweischneidige Natur von KI als Werkzeug **für beispiellose Chancen und komplexe Risiken.**
Da sich die Leistungsfähigkeit und Verbreitung von KI-Systemen weiterentwickelt, bieten die verantwortungsvollen Strategien von Anthropic wertvolle Einblicke in die ethische Entwicklung und die gesellschaftlichen Auswirkungen.
Das Gleichgewicht zwischen Innovation und Verantwortlichkeit und Weitsicht bleibt von entscheidender Bedeutung, da die Menschheit immer tiefer in eine Ära vordringt, die von intelligenten Maschinen geprägt ist.
Durch die Betonung von Grundwerten und das Eintreten für kooperative globale Standards bekräftigt Amodei die Vorstellung, dass KI zwar das Potenzial für immense Fortschritte birgt, wir aber wachsam bleiben müssen, wenn es darum geht, ihren Kurs zu steuern.

Nachfolgend das unkorrigiertes Transkript des US-Videos, das ich automatisiert ins Deutsche habe übertragen lassen:

FROMAN: Nun, guten Abend, allerseits. Willkommen. Mein Name ist Mike Froman. Ich bin Präsident des Rates. Und es ist uns eine grosse Freude, Sie heute Abend hier zu einer unserer **CFR (Councel on Foreign Relations)** **CEO Speaker Series** zu haben und den **CEO und Mitbegründer von Anthropic** Dario Amodei heute Abend bei uns zu haben.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Dario war Vice-President of Research bei OpenAI, wo er an der Entwicklung von GPT-2 und -3 beteiligt war. Und bevor er zu OpenAI kam, arbeitete er bei **Google Brain als Senior Research Scientist**. Ich werde etwa dreissig Minuten mit Dario sprechen. Dann öffnen wir es für Fragen von Leuten hier im Saal. Wir haben hier etwa 150 Leute. Wir haben etwa 350 online. Und so werden wir versuchen, auch einige ihrer Fragen einzubringen. Willkommen.

AMODEI: Danke, dass Sie mich eingeladen haben.

FROMAN: **Sie haben also OpenAI verlassen, um Anthropic zu gründen, ein gemeinnütziges Unternehmen, das sich an erster Stelle befindet**. Warum gehen? Was sind die Kernwerte von Anthropic? Und wie manifestieren sie sich in Ihrer Arbeit? Und lassen Sie mich einfach sagen, ein Zyniker würde sagen, nun, diese Mission zuerst, das ist alles Marketing. **Können Sie uns einige konkrete Beispiele dafür geben, wie Ihr Produkt und Ihre Strategie Ihre Mission widerspiegeln?**

AMODEI: Also, ja, wenn ich ... wenn ich würde, wissen Sie, dann würde ich einfach zurückgehen und den Kontext irgendwie festlegen. Wissen Sie, wir sind Ende 2020 gegangen. Ich denke, in den Jahren 2019 und 2020 ist etwas passiert, das ich und eine Gruppe innerhalb von OpenAI, die schliesslich zu meinen Mitbegründern bei Anthropic wurden, glaube ich, zu den ersten gehörten, **die es erkannt haben. Man nennt sie eine Art Skalierungsgesetze oder heute die Skalierungshypothese**. Und die Grundhypothese ist einfach. Es besagt – und es ist eine wirklich bemerkenswerte Sache, und ich kann nicht genug betonen, wie unwahrscheinlich es damals schien – wenn man mehr Berechnungen und mehr Daten benötigt, um KI-Systeme mit relativ einfachen Algorithmen zu trainieren, werden sie auf der ganzen Linie bei allen Arten von kognitiven Aufgaben besser. Und wir haben diese Trends damals gemessen, als das Training von Modellen 1.000 oder 10.000 US-Dollar kostete. Das ist also eine Art Budget für akademische Zuschüsse.
Und wir prognostizieren, dass sich diese Trends fortsetzen werden, selbst wenn das Training von Modellen 100 Millionen, eine Milliarde, 10 Milliarden Dollar kostet, und das sind wir jetzt. Und in der Tat, dass, wenn die Qualität der Modelle und ihr Intelligenzniveau anhielten, sie enorme Auswirkungen auf die Wirtschaft haben würden. **Es war sogar das erste Mal, dass wir erkannten, dass sie wahrscheinlich sehr ernste Auswirkungen auf die nationale Sicherheit haben würden.**

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Wir hatten im Allgemeinen das Gefühl, dass die Führung von OpenAI mit dieser allgemeinen Skalierungshypothese einverstanden war, obwohl viele Leute innerhalb und ausserhalb dies nicht taten.

Aber die zweite Erkenntnis, die wir hatten, war, dass wir, wenn die Technologie ein solches Mass an Bedeutung haben sollte, wirklich gute Arbeit leisten mussten, um sie zu entwickeln. Wir mussten es wirklich richtig machen.
Insbesondere sind diese Modelle einerseits sehr unvorhersehbar.
Sie sind von Natur aus statistische Systeme.
Eine Sache, die ich oft sage, ist, dass wir sie mehr anbauen, als wir sie bauen. Sie sind wie das Gehirn eines Kindes, das sich entwickelt.
Sie zu kontrollieren, sie zuverlässig zu machen, ist also sehr schwierig.
Der Prozess, sie zu trainieren, ist nicht einfach.
Allein aus Sicht der Systemsicherheit ist es also sehr wichtig, diese Dinge vorhersehbar und sicher zu machen.
Und dann gibt es natürlich noch die Verwendung von ihnen – **die Nutzung durch Menschen, die Nutzung durch Nationalstaaten, die Wirkung, die sie haben, wenn Unternehmen sie einsetzen.**
Und so hatten wir wirklich das Gefühl, dass wir etwas aufbauen mussten. Diese Technologie ist, wissen Sie, absolut richtig. Wissen Sie, OpenAI, ein bisschen, wie Sie angedeutet haben, wurde mit einigen Behauptungen gegründet, dass sie genau das tun würden. Aber aus einer Reihe von Gründen, auf die ich nicht im Detail eingehen werde, hatten wir nicht das Gefühl, dass die Führung dort diese Dinge ernst nahm. Und so beschlossen wir, loszuziehen und das auf eigene Faust zu machen. Und die letzten vier Jahre waren eigentlich eine Art Experiment fast Seite an Seite, wissen Sie, was passiert, wenn man versucht, Dinge auf die eine Weise zu tun, und was passiert, wenn man versucht, Dinge auf die andere Weise zu tun, und wie es sich entwickelt hat. Also, wissen Sie, ich werde ein paar Beispiele dafür geben, wie wir, wissen Sie, wirklich, glaube ich, ein Engagement für diese Ideen gezeigt haben.
Eine davon ist, dass wir sehr früh in die Wissenschaft der sogenannten **mechanistischen Interpretierbarkeit investiert haben**, d. h. in die KI-Modelle hineinschauen und versuchen genau zu verstehen, warum sie tun, was sie tun. Einer unserer sieben Mitbegründer, Chris Olah, ist der Begründer des Bereichs der mechanistischen Interpretierbarkeit. **Das hatte keinen kommerziellen Wert, oder zumindest keinen kommerziellen Wert in den ersten vier Jahren,** in denen wir daran gearbeitet haben. Es fängt gerade erst an, ein bisschen in der Ferne zu sein.

Nichtsdestotrotz hatten wir ein Team, das die ganze Zeit daran gearbeitet hat, und das in einem harten kommerziellen Wettbewerb, weil wir glauben, dass das Verständnis, was in diesen Modellen vor sich geht, ein öffentliches Gut ist, das allen zugutekommt. **Und wir haben unsere gesamte Arbeit dazu veröffentlicht, damit auch andere davon profitieren können**. Wissen Sie, ich denke, ein weiteres Beispiel ist, dass wir auf die Idee der konstitutionellen KI gekommen sind, bei der KI-Systeme so trainiert werden, dass sie einer Reihe von Prinzipien folgen. Anstatt sie mit Daten oder Massendaten oder menschlichem Feedback zu trainieren. Wissen Sie, das ermöglicht es Ihnen, aufzustehen, sagen Sie, Sie wissen schon, vor dem Kongress, und zu sagen: Das sind die Prinzipien, nach denen wir unser Modell trainiert haben.

Als wir zum ersten Mal darauf kamen – wissen Sie, als wir unser erstes Produkt hatten, unsere – Sie wissen schon, **unsere erste Version von Claude**, die unser Modell ist, haben wir die Veröffentlichung dieses Modells tatsächlich um etwa sechs Monate verschoben, weil es sich um eine so neue Technologie handelte, dass wir uns über die Sicherheitseigenschaften nicht sicher waren. Wir waren uns nicht sicher, ob wir diejenigen sein wollten, die ein Rennen starten. **Das war kurz vor ChatGPT**. Also, wissen Sie, wir hatten wohl die Gelegenheit, den ChatGPT-Moment zu nutzen, und wir haben uns entschieden, etwas später zu veröffentlichen. Was meiner Meinung nach echte kommerzielle Konsequenzen hatte, aber die Kultur des Unternehmens bestimmte. Ein letztes Beispiel, das ich nennen möchte, ist, **dass wir die ersten waren, die eine sogenannte verantwortungsvolle Skalierungsrichtlinie hatten**. Dies misst also die Risikokategorien von Modellen, während sie skalieren. **Und wir müssen immer strengere Sicherheitsmassnahmen ergreifen**, wenn wir diese Punkte erreichen.
Und so waren wir die ersten, die das veröffentlicht haben, die ersten, die sich dazu verpflichtet haben. Und dann ein paar Monate – innerhalb weniger Monate, nachdem wir das getan hatten, folgten alle anderen Unternehmen diesem Beispiel. Und so konnten wir ein Zeichen für das Ökosystem setzen. Und wenn ich mir anschaue, was die anderen Unternehmen getan haben, haben wir bei diesen Themen **oft eine Vorreiterrolle übernommen** und oft die anderen Unternehmen dazu gebracht, uns zu folgen. Nicht immer. Manchmal machen sie etwas Grossartiges und wir folgen ihnen. Aber ich denke, es gab eine gute – es gab eine gute Geschichte, in der wir uns an unsere Verpflichtungen gehalten haben. Und wissen Sie, ich würde das mit dem vergleichen, was wir von einigen der anderen Unternehmen in ihrem Verhalten gesehen haben.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Wir haben jetzt eine mehrjährige Geschichte. Und wissen Sie, ich drücke die Daumen, dass unsere Verpflichtungen bisher ziemlich gut gehalten haben.

FROMAN: **Ich möchte sowohl über die Risiken als auch über die Chancen sprechen**, die Sie im Zusammenhang mit KI genannt haben. Aber da Sie das Problem der verantwortungsvollen Skalierung angesprochen haben, kommen wir darauf zurück. Wir sind jetzt auf Stufe zwei. AMODEI: Ja, also...

FROMAN: Und auf welcher Ebene ist es existenziell? Woher wissen wir, wann wir Level drei erreicht haben? Und wenn du Level drei erreichst, kannst du dann rückwärts gehen, oder wird es nur noch schlimmer?

AMODEI: Ja. Die Art und Weise, wie unsere Politik der verantwortungsvollen Skalierung aufgebaut ist, ist also im Grunde genommen – wissen Sie, und die **Analogie bezog sich auf das Niveau der biologischen Sicherheit**. Also, wissen Sie, **das Biosicherheitssystem ist, wissen Sie, wie gefährlich verschiedene Krankheitserreger sind.** Und so sagten wir, lasst uns ein KI-Sicherheitsniveau haben. Und so ist KI-Sicherheitsstufe zwei ein Niveau, auf dem wir uns derzeit befinden. Und das sind Systeme, die leistungsfähig sind, aber die Risiken, die sie darstellen, sind vergleichbar mit den Risiken, die andere Arten von Technologien mit sich bringen.
ASL-3, von dem ich glaube, dass sich unsere Modelle langsam annähern – das letzte Modell, das wir veröffentlicht haben, haben wir gesagt, dass dieses Modell noch nicht ASL-3 ist, aber es ist auf dem Weg dorthin. ASL-3 ist charakterisiert – **und wir konzentrieren uns sehr stark auf die nationale Sicherheit**. Sehr ernste Risiken, die in keinem Verhältnis zu den Risiken stehen, die normale Technologien mit sich bringen.
Ein ASL-3-Modell ist also so konzipiert, dass es Ihnen ermöglichen könnte, in den Bereichen chemische, biologische oder radiologische Waffen einer ungelernten Person zu ermöglichen, einfach durch das Gespräch mit dem Modell und das Befolgen seiner Anweisungen Dinge zu tun, für die Sie heute beispielsweise einen Doktortitel in Virologie haben müssten.

Wenn das also möglich ist, wenn diese Risiken nicht gemindert werden, dann würde das die Zahl der Menschen auf der Welt erhöhen, die in der Lage sind, diese hochgradig zerstörerischen Dinge zu tun, von, sagen wir, Zehntausenden heute, auf Dutzende von Millionen, sobald die Modelle verfügbar sind.

Und wenn die Modelle dazu in der Lage sind, müssen wir Abhilfemassnahmen ergreifen, damit die Modelle nicht bereit sind, diese Informationen zur Verfügung zu stellen, und Sicherheitseinschränkungen, damit die Modelle nicht gestohlen werden.
Und wissen Sie, ich denke, wir nähern uns dem.
Das könnten wir dieses Jahr tatsächlich erreichen. Und wir glauben, dass wir eine Geschichte dafür haben, wie wir diese Art von Modellen sicher einsetzen können, **indem wir ihnen die Fähigkeit nehmen, diesen sehr engen Bereich gefährlicher Aufgaben zu erledigen, ohne ihre kommerzielle Rentabilität zu beeinträchtigen.**

FROMAN: Das ist also ein ziemlich enger Satz von Aufgaben. Wie Sie sagen, werden Sie das Modell einfach daran hindern, diese Fragen zu beantworten.

AMODEI: **Ja, verhindern Sie, dass das Modell sich an dieser Art von Aufgaben beteiligt. Das ist nicht einfach, oder?** Man kann sagen, ich besuche einen Virologiekurs an der Stanford University. Ich arbeite an meinen Kursarbeiten, z.B. können Sie mir sagen, wie man dieses spezielle Plasmid herstellt?
Und das Model muss schlau genug sein, um nicht darauf hereinzufallen und zu sagen: Hey, weisst du, das ist eigentlich nicht die Art von Dingen, die man fragen würde...

FROMAN: **Sie klingen wie ein Bioterrorist**. Ich werde Ihre Frage nicht beantworten.

AMODEI: Du klingst, als hättest du böse Absichten.

FROMAN: Ja. Aber es ist irgendwie auf deine eigene Vorstellungskraft beschränkt, oder auf unsere eigene Vorstellungskraft, was all das Schauspiel sein könnte. Es gibt viele Dinge, die wir über diese vier Kategorien hinaus vielleicht nicht vorhersehen.

AMODEI: Ja. ja. Ich meine, wissen Sie, ich denke, das ist ein Problem, das – so wie jedes Mal, wenn wir ein neues Modell auf den Markt bringen, positive Anwendungen dafür gibt, die die Leute finden, die wir nicht erwartet haben. Ich gehe davon aus, dass es auch negative Bewerbungen geben wird.
Wir überwachen die Modelle immer für verschiedene Anwendungsfälle, um dies zu entdecken, damit wir einen kontinuierlichen Prozess haben, bei dem wir nicht überrascht werden.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Wenn Sie wissen, dass wir uns Sorgen machen, dass jemand etwas Böses mit dem sechsten Modell anstellen wird, sind hoffentlich einige frühe Anzeichen dafür im fünften Modell zu sehen. Und wir überwachen es. Aber das ist das grundlegende Problem der Modelle. **Man weiss nicht wirklich, wozu sie fähig sind**. **Man weiss nicht wirklich, wozu sie fähig sind, bis sie bei einer Million Menschen eingesetzt werden**. Sie können im Voraus testen. Sie können – wissen Sie, Sie können, Sie wissen, Sie können Ihre Forscher dazu bringen, gegen sie zu wettern. Sie können, wissen Sie, sogar die Regierung haben – **wir arbeiten mit der Regierung zusammen,**
AISIs testen sie. **Aber die harte Wahrheit ist, dass es keine Möglichkeit gibt, sicher zu sein. Sie sind nicht wie Code, bei dem Sie eine formale Verifizierung durchführen können**. **Was sie tun können, ist unvorhersehbar**. Es ist einfach so, wenn ich an dich oder mich denke, anstatt an das Model. Wissen Sie, wenn ich der Qualitätssicherungsingenieur für mich oder Sie bin, kann ich dann garantieren, dass es eine bestimmte Art von schlechtem Verhalten gibt, zu dem Sie logischerweise nicht fähig sind, das nie passieren wird? Menschen arbeiten nicht auf diese Weise.

FROMAN: Lassen Sie uns über die Chancen sprechen, die Aufwärtschancen.

AMODEI: Absolut.

FROMAN: Ende letzten Jahres haben Sie einen Essay geschrieben, **"Machines of Loving Grace",** in dem Sie über einige der Vorteile sprechen. Wie man in der Biologie den Fortschritt eines Jahrzehnts erreichen könnte,
zum Beispiel in einem Jahr, wie die Maschinen so schlau sein würden wie alle Nobelpreisträger, was wahrscheinlich einige von ihnen deprimiert.
Nennen Sie uns die Vorteile.
Nennen Sie uns Ihr Best-Case-Szenario, was KI produzieren wird.

AMODEI: Ja. Also gehe ich zurück, indem ich mit dem Exponential beginne. Wenn wir ins Jahr 2019 zurückgehen, waren die Modelle kaum in der Lage, einen zusammenhängenden Satz oder einen zusammenhängenden Absatz zu geben. Leute wie ich dachten natürlich, dass das eine erstaunliche Leistung ist, zu der Models nicht in der Lage sind. **Und wissen Sie, wir hatten diese Vorhersagen, dass die Modelle in fünf Jahren Milliarden von Dollar an Einnahmen generieren werden. Sie werden uns beim Programmieren helfen.**

Wir können mit ihnen sprechen, als wären sie – als wären sie Menschen. Sie werden so viel wissen wie Menschen.

Und es gab all diese prinzipienlosen Einwände, warum das nicht passieren konnte.
Wissen Sie, die gleichen exponentiellen Trends, die gleichen Argumente, die das vorhergesagt haben, sagen voraus, dass wir, wenn wir weitere zwei, drei Jahre, vielleicht vier Jahre weitermachen, zu all dem Kommen werden.
Wir werden zu Modellen gelangen, die so intelligent sind wie Nobelpreisträger in einer ganzen Reihe von Bereichen.
Sie werden nicht nur mit ihnen chatten, sie werden in der Lage sein, alles zu tun, was Sie auf einem Computer tun können.

Im Grunde genommen jede Remote-Arbeit, die Menschen machen, jede Modalität, in der Lage zu sein, Aufgaben zu erledigen, die Tage, Wochen, Monate dauern. Die Art von evokativer Formulierung, die ich dafür in "Machines of Loving Grace" verwendet habe, war, als hätte man ein Land voller Genies in einem Rechenzentrum.

Wie ein Land der genialen Remote-Arbeiter, die nicht alles können, oder?
Es gibt Einschränkungen in der physischen Welt. Und ich denke, das klingt für viele Leute immer noch verrückt, aber schauen Sie auf frühere exponentielle Trends zurück.

Schauen Sie sich die Anfänge des Internets an und wie wild die Vorhersage schien und was tatsächlich eingetreten ist. Da bin ich mir nicht sicher. Ich würde sagen, ich habe vielleicht 70 oder 80 Prozent Selbstvertrauen. Wissen Sie, es könnte sehr gut sein, dass die Technologie dort aufhört, wo sie ist, oder in ein paar Monaten aufhört, und, wissen Sie, die Essays, die ich geschrieben habe, und Dinge, die ich gesagt habe, wissen Sie, bei Ereignissen wie diesem, die Leute verbringen die nächsten zehn Jahre damit, mich auszulachen.
Aber das wäre nicht meine Wette.

FROMAN: Lassen Sie uns einfach darauf aufbauen, denn über das **Thema Arbeitsplätze und die Auswirkungen, die KI wahrscheinlich auf die Beschäftigung haben wird, gibt es eine ziemlich grosse Debatte.**
Wo befinden Sie sich auf dem Spektrum – bevor ich dazu komme – wie lange wird es dauern, **bis KI, sagen wir, den Chef eines Think Tanks ersetzt?**
Ich frage nach einem Freund. (Gelächter.) Eigentlich wie – dazu kommen wir.
Und wo befindest du dich im Spektrum, in dem jeder in der Lage sein wird, einige wirklich coole Dinge zu tun, und sie werden in der Lage sein, so viel mehr Dinge zu tun, als sie jetzt können, im Gegensatz zu jedem, der auf seinem Sofa sitzt und BGE sammelt?

AMODEI: Ja. Ich denke also, dass es eine wirklich komplizierte Mischung aus diesen beiden Dingen sein wird, die auch von den politischen Entscheidungen abhängt, die wir treffen.

FROMAN: Sie können auch die Frage nach dem Think-Tank beantworten, wenn Sie möchten, aber – (Gelächter).

AMODEI: Ja. Also, ich meine, ich habe es wohl nicht getan – ich habe meine Antwort auf die letzte Frage beendet, ohne all die grossartigen Dinge zu sagen, die passieren werden.
Ehrlich gesagt, das, was mich am optimistischsten stimmt, bevor ich zu Jobs komme, sind Dinge in den Biowissenschaften – Biologie, Gesundheit, Neurowissenschaften.
Wissen Sie, ich denke, wenn wir uns ansehen, was in den letzten hundert Jahren in der Biologie passiert ist, dann sind das, was wir gelöst haben, einfache Krankheiten.
Die Lösung von viralen und bakteriellen Krankheiten ist eigentlich relativ einfach, denn es ist das Äquivalent zur Abwehr eines fremden Eindringlings in Ihrem Körper.
Der Umgang mit Dingen wie Krebs, Alzheimer, Schizophrenie, schwerer Depression – das sind Krankheiten auf Systemebene.
Wenn wir diese Probleme unabhängig von der Art der beruflichen Situation mit KI auf einer Grundlage lösen können, werden wir eine viel bessere Welt haben. Und ich denke, wir werden sogar – wenn wir uns der Seite der psychischen Erkrankungen zuwenden – eine Welt haben, in der es für die Menschen zumindest einfacher ist, einen Sinn zu finden. Ich bin also sehr optimistisch, was das angeht.
Aber jetzt, wo ich mich der beruflichen Seite ankomme, habe ich eine ziemliche Sorge darüber. Auf der einen Seite denke ich, dass der komparative Vorteil ein sehr mächtiges Instrument ist.
Wenn ich mir das Programmieren anschaue, das ist ein Bereich, in dem KI die grössten Fortschritte macht, dann stellen wir fest, dass wir nicht mehr weit von der Welt entfernt sind – ich denke, wir werden in drei bis sechs Monaten dort sein –, in der KI 90 Prozent des Codes schreibt.

Und in zwölf Monaten könnten wir dann in einer Welt sein, in der KI im Wesentlichen den gesamten Code schreibt.

Aber der Programmierer muss immer noch spezifizieren, was sind – was sind **die Bedingungen für das, was Sie tun**, was – wissen Sie, was ist die gesamte App, die Sie zu erstellen versuchen, was ist die allgemeine Designentscheidung?
Wie arbeiten wir mit anderem Code zusammen, der geschrieben wurde? Wissen Sie, wie wir einen gesunden Menschenverstand haben, ob es sich um ein sicheres oder ein unsicheres Design handelt? Solange es also diese kleinen Teile gibt, die ein Programmierer, ein menschlicher Programmierer, tun muss, in denen die KI nicht gut ist, denke ich, dass die menschliche Produktivität tatsächlich gesteigert wird.
Aber auf der anderen Seite denke ich, dass all diese kleinen Inseln irgendwann von KI-Systemen abgegriffen werden.
Und dann werden wir irgendwann den Punkt erreichen, an dem die KIs alles können, was Menschen können.

Und ich denke, das wird in jeder Branche passieren.

Ich denke, es ist eigentlich besser, dass es uns allen passiert, als dass es passiert – wissen Sie, dass es die Leute zufällig auswählt.
Ich denke tatsächlich, dass das gesellschaftlich spaltendste Ergebnis darin besteht, wenn zufällig 50 Prozent der Jobs plötzlich von KI erledigt werden, denn das bedeutet, dass die gesellschaftliche Botschaft lautet:
Wir wählen die Hälfte aus – wir wählen wahllos die Hälfte der Menschen aus und sagen: Du bist nutzlos, du wirst abgewertet, du bist unnötig.

FROMAN: **Und stattdessen werden wir sagen, ihr seid alle nutzlos?**
(Gelächter.)

AMODEI: **Nun, wir werden alle dieses Gespräch führen müssen, oder?**
Wir werden – wir werden – wir werden – wir müssen uns ansehen, was technologisch möglich ist, und sagen, **wir müssen über Nützlichkeit und Nutzlosigkeit auf eine andere Art und Weise nachdenken als zuvor, richtig? Unsere derzeitige Denkweise war nicht haltbar**.
Ich weiss nicht, was die Lösung ist, aber es muss so sein – es muss anders sein als, wir sind alle nutzlos, oder? Wir sind alle nutzlos ist eine nihilistische Antwort. **Mit dieser Antwort werden wir nirgendwo hinkommen**.
Wir werden uns etwas anderes einfallen lassen müssen.

FROMAN: Das ist kein sehr optimistisches Bild. (Gelächter.) Ist es, was es ist?

AMODEI: Ich würde das tatsächlich in Frage stellen. Weisst du, ich denke über viele der Dinge nach, die ich tue – weisst du, ich verbringe viel Zeit, zum Beispiel mit Schwimmen.
Ich verbringe Zeit damit, Videospiele zu spielen. Ich betrachte menschliche Schachchampions.
Weisst du, als Deep Blue Kasparov besiegte, und das war vor fast dreissig Jahren, könnte man meinen, dass Schach danach als eine sinnlose Aktivität angesehen werden würde. Doch genau das Gegenteil ist eingetreten. Menschliche Schachmeister wie Magnus Carlsen sind Berühmtheiten.
Ich glaube, er ist sogar, so etwas wie ein Model. Er ist so ein Held. Ich denke also, dass es etwas gibt, wo wir etwas aufbauen können – wir können eine Welt schaffen, in der das menschliche Leben sinnvoll ist und Menschen, vielleicht mit Hilfe von KIs, vielleicht in Zusammenarbeit mit KIs, wirklich grossartige Dinge bauen. Das bin ich also nicht – so pessimistisch bin ich eigentlich nicht. Aber wenn wir es falsch handhaben, gibt es meiner Meinung nach vielleicht nicht so viel Spielraum für Fehler.

FROMAN: **Vor ein paar Monaten wurde DeepSeek veröffentlicht**. In dieser Stadt herrschte ein gewisses Mass an Panik, würde ich sagen, deswegen.
Die Leute sprachen von einem Sputnik-Moment. War es ein Sputnik-Moment? Und was sagt uns das darüber, ob die Skalierungsregeln, die Sie festgelegt haben, mehr Rechenleistung, mehr Daten, bessere Algorithmen benötigen, ob diese Regeln immer noch gelten oder ob es einige Abkürzungen gibt.

AMODEI: Ja. Ich denke, DeepSeek war es also tatsächlich – anstatt die Skalierungsgesetze zu widerlegen, **denke ich, dass DeepSeek tatsächlich ein Beispiel für die Skalierungsgesetze war**.
Es gibt also zwei Dynamiken – ich hatte einen Beitrag darüber –, aber
zwei Dynamiken finden gleichzeitig statt.
Eine davon ist, dass die Kosten für die Erzeugung eines bestimmten Niveaus an Modellintelligenz sinken, etwa um das 4-fache pro Jahr. Das liegt daran, dass wir immer besser darin werden, algorithmisch die gleichen Ergebnisse mit weniger Kosten zu erzielen. Mit anderen Worten, wir verschieben die Kurve. Sie können für – wissen Sie, ein Jahr später können Sie – Sie wissen schon, ein so gutes Modell bekommen, wie Sie vor einem Jahr bekommen konnten, wenn Sie das 4-fache ausgegeben haben. Sie können ein 4X besseres Modell erhalten, indem Sie den gleichen Betrag ausgeben. Aber was das ökonomisch bedeutet, ist, dass, unabhängig vom wirtschaftlichen Wert des aktuellen Modells einer bestimmten Intelligenz, die Tatsache, dass man es 4x billiger machen kann, bedeutet, dass wir viel mehr davon machen, und tatsächlich

einen zusätzlichen Anreiz bietet, mehr Geld auszugeben, um intelligentere Modelle zu produzieren, die einen höheren wirtschaftlichen Wert haben. Und selbst wenn die Kosten für die Produktion eines bestimmten Niveaus an Intelligenz gesunken sind, ist der Betrag, den wir bereit sind auszugeben, gestiegen. Tatsächlich ist es schnell gestiegen, etwa das 10-fache pro Jahr, trotz dieses 4-fachen Anstiegs pro Jahr, oder?

Das ist aufgefressen worden und noch mehr von der gerechten Gesellschaft, die Wirtschaft will mehr Intelligenz. Sie will intelligentere Modelle. Das ist also eine Art Hintergrund für DeepSeek. Und DeepSeek war buchstäblich nur ein weiterer Datenpunkt auf der Kostensenkungskurve. Es war nichts Ungewöhnliches.

Es war nicht so, dass diese US-Unternehmen Milliarden ausgeben und DeepSeek es für ein paar Millionen getan hat.

Die Kosten waren nicht aus der Reihe. Sie haben ein paar Millionen für das Modell ausgegeben. Was US-Unternehmen ausgeben, steht nicht im Widerspruch dazu.

Sie haben, wie wir, Milliarden für all die Forschung und Entwicklung und den Aufwand rund um das Modell ausgegeben. Wenn man sich anschaut, wie viele Chips sie haben, ist es ungefähr gleichauf. Nun, ich denke, es ist besorgniserregend, denn bis vor kurzem gab es nur drei, vier, vielleicht fünf Unternehmen, die Teil dieser Kurve waren und Frontier-Modelle produzieren konnten. Und sie waren alle in den USA.

DeepSeek ist das erste Mal – was wirklich bemerkenswert ist – es ist das erste Mal, dass ein Unternehmen in China in der Lage ist, die gleiche Art von technischen Innovationen zu produzieren wie Unternehmen wie Anthropic oder OpenAI oder Google. Das ist in der Tat sehr wichtig. Und das macht mir tatsächlich Sorgen.

FROMAN: Nun, einige argumentieren, dass das Aufkommen von DeepSeek **bedeutet, dass Exportkontrollen nicht funktionieren, nicht funktionieren können, wir sollten aufhören zu versuchen, den Export unserer fortschrittlichsten Chips zu kontrollieren.** Andere sagen, das bedeute, dass wir die Exportkontrollen verdoppeln sollten. Wie stehen Sie dazu?

AMODEI: Ja. Also, wissen Sie, ich denke, es ist eine Implikation des Rahmens, den ich gerade gegeben habe, **dass die Exportkontrollen tatsächlich ziemlich wichtig sind, denn ja, es gibt diese Kostensenkungskurve, aber an jedem Punkt entlang der Kurve, egal wie stark die Kurve verschoben ist, ist es immer so, dass je mehr Chips man ausgibt, desto mehr Geld gibt man aus.**

Das bessere Modell, das Sie bekommen, oder?
Wenn es so ist, dass ich früher eine Milliarde Dollar ausgeben und ein Modell bekommen konnte, das in Ordnung war, kann ich jetzt eine Milliarde Dollar ausgeben und ein Modell bekommen, das viel besser ist, und ich kann ein OK-Modell für 10 Millionen Dollar bekommen. Das bedeutet nicht, dass die Exportkontrollen versagt haben. Das bedeutet, dass es zu einer Sache mit höherem Einsatz geworden ist, Ihre Gegner daran zu hindern, ein Milliarden-Dollar-Modell zu bekommen, weil Sie ein intelligenteres Modell für eine Milliarde Dollar bekommen können. Und ja, DeepSeek war – wissen Sie, sie hatten eine relativ kleine Menge an Rechenleistung, bestehend aus Chips, die die Exportkontrollen umgingen, und einigen Chips, die geschmuggelt wurden. Aber ich denke, wir steuern auf eine Welt zu, in der wir, OpenAI und Google, Milliarden bauen, vielleicht in Dutzenden von Millionen Chips, die Dutzende von Milliarden Dollar oder mehr kosten. Es ist sehr schwer, dass das geschmuggelt wird. Wenn wir Exportkontrollen einführen, können wir das in China vielleicht sogar verhindern. Wenn wir dagegen – wenn wir es nicht tun, denke ich, dass sie mit uns gleichgestellt sein könnten. Und so war ich ein grosser Befürworter der Diffusionsregel. Ich bin seit mehreren Jahren ein grosser Befürworter von Exportkontrollen, noch bevor DeepSeek herauskam, weil wir diese Dynamik kommen sahen. Und deshalb denke ich, dass es tatsächlich eines der wichtigsten Dinge ist, nicht nur in der KI, sondern in allen Bereichen, für die nationale Sicherheit der Vereinigten Staaten, dass wir verhindern, dass China Millionen dieser sehr leistungsstarken Chips bekommt.

FROMAN: Die Diffusionsregel, so wie ich sie verstehe, hat die Welt gespalten – das ist eine EO der Biden-Regierung –,
die die Welt in drei Lager eingeteilt hat,
wer Zugang zu was bekommen kann, in Bezug auf Chips von uns. Einige befürchten, dass die Länder, die nicht in der Spitzengruppe stehen, nur von China bedient werden und dass China am Ende die KI-Infrastruktur für die überwiegende Mehrheit der Welt betreiben wird. Das ist Ihnen passiert?

AMODEI: Ja. Mein Verständnis der Diffusionsregel ist, und, wissen Sie, ich verstehe, dass die neue Regierung sich das ansieht, aber es gibt viele Teile, mit denen sie sympathisieren. Die Art und Weise, wie es die Dinge tatsächlich aufbaut, sind diese Tier-2-Länder. Die Tier-1-Länder sind also, wie die Mehrheit der entwickelten Welt.

FROMAN: Aber nicht alle.

AMODEI: **Nicht alle. Stufe drei sind, wissen Sie, eingeschränkte Länder wie China oder Russland. Stufe zwei sind, wissen Sie, Länder in der Mitte. Tatsächlich kann man in diesen Ländern eine sehr grosse Anzahl von Chips haben, wenn die Unternehmen, die sie hosten, in der Lage sind, eidesstattliche Erklärungen und Garantien abzugeben, die im Grunde besagen, dass wir keine Tarnfirma für China sind. Wir versenden die Rechenleistung oder das, was mit der Rechenleistung gemacht wird, nicht nach China. Und so gibt es wirklich die Möglichkeit, eine Menge US-Chips, eine Menge US-Infrastruktur in diesen Ländern zu bauen, solange sie die Sicherheitsbeschränkungen einhalten**.

Ich denke, der zweite Teil davon ist, ja, theoretisch könnten Unternehmen auf die Verwendung chinesischer Chips umsteigen. Aber chinesische Chips sind eigentlich ziemlich minderwertig. Nvidia ist weit vor Huawei, dem Hauptproduzenten von Chips für China. **Etwa vier Jahre im Voraus. Ich denke, diese Lücke wird sich irgendwann schliessen, im Laufe von, ich weiss nicht, zehn oder zwanzig Jahren.**
Wahrscheinlich könnten die Exportkontrollen sogar die Wirkung haben, China zu stimulieren. Aber der Tech-Stack ist so tief. Und ich denke, dass die nächsten zehn Jahre, in denen wir bei der Hardware stark voraus sein werden, tatsächlich die kritische Phase für die Etablierung einer Dominanz in dieser Technologie sind, die, wie ich behaupten würde,
wer auch immer eine Dominanz in dieser Technologie etabliert, überall militärische und wirtschaftliche Dominanz haben wird.

FROMAN: Die letzte Regierung hat einen Dialog mit China über KI aufgenommen. Welche Aussichten gibt es für einen solchen Dialog?
Wo könnten wir uns mit China einigen?
Und legen sie Wert auf verantwortungsvolle Skalierung?

AMODEI: Ja. Also, wissen Sie, ich würde mich selbst beschreiben – und natürlich war ich nicht Teil eines dieser Gespräche, aber ich habe ein wenig davon gehört – ich würde mich als Unterstützer dieses Dialogs beschreiben, aber nicht besonders optimistisch, dass es funktionieren würde.

Die Technologie hat also ein so grosses wirtschaftliches und militärisches Potenzial, dass man sich vorstellen kann, zwischen Unternehmen in den USA oder unseren demokratischen Verbündeten Gesetze zu verabschieden, die eine gewisse Zurückhaltung schaffen.

Wenn es nur so ist, als ob zwei Seiten darum wetteifern, diese Technologie zu entwickeln, die so viel wirtschaftlichen und militärischen Wert hat, vielleicht mehr als alles andere zusammen, **ist es schwer vorstellbar, dass sie sich deutlich verlangsamen. Ich denke, da gibt es ein paar Dinge. Eines ist das Risiko, dass die KI-Modelle autonom auf eine Weise handeln, die nicht im Einklang mit den menschlichen Interessen steht, oder?**
Wenn Sie ein Land voller Genies in einem Rechenzentrum haben, stellt sich die Frage natürlich: Wie könnten Sie diese Frage nicht stellen – nun, was ist ihre Absicht?
Was haben sie vor? Sie würden sicherlich fragen, ob sie jemand kontrolliert? Handeln sie im Namen von jemandem? Aber Sie würden auch fragen, was ist ihre Absicht? Und weil wir diese Systeme weiterentwickeln, sie nicht trainieren, glaube ich nicht, dass man davon ausgehen kann, dass sie genau das tun, was ihre menschlichen Designer oder Benutzer von ihnen erwarten.
Ich denke also, dass das ein reales Risiko darstellt. Ich denke, es könnte eine Bedrohung für die gesamte Menschheit sein. Und wie bei Fragen der nuklearen Sicherheit oder der nuklearen Proliferation gibt es wahrscheinlich eine gewisse Möglichkeit, begrenzte Massnahmen zu ergreifen, um diesem Risiko zu begegnen.
Ich bin also relativ optimistisch, dass vielleicht etwas Enges gemacht werden könnte. Je stärker die Beweise dafür sind, dass das kommen wird – wissen Sie, **im Moment ist das eine Art spekulative Sache.**
Aber wenn es starke Beweise dafür gibt, dass dies unmittelbar bevorsteht, dann wäre vielleicht eine weitere Zusammenarbeit mit China möglich.
Ich bin zuversichtlich, dass wir versuchen können, etwas in diesem Bereich zu tun, aber ich glaube nicht, dass wir die Dynamik ändern werden nationalen Wettbewerbs zwischen den beiden.

FROMAN: Letzte Frage, bevor wir es öffnen. Sie haben vor kurzem, glaube ich, OSTP einen Aktionsplan vorgelegt – **einen vorgeschlagenen Aktionsplan für die neue Regierung,** was sie in diesem Bereich tun sollte.
Was sind die Hauptelemente dieses Plans?

AMODEI: Ja. Ich denke also, **dass es drei Elemente gibt**, die mit der Art der Sicherheit und der nationalen Sicherheit zu tun haben, und drei Elemente mit Chancen.
Der erste ist das, worüber wir gesprochen haben, nämlich sicherzustellen, dass wir diese Exportkontrollen beibehalten. Ich glaube wirklich, dass dies – in allen Bereichen, nicht nur in Bezug auf KI – **die wichtigste Politik für die nationale Sicherheit der Vereinigten Staaten ist**.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Die zweite Sache ist etwas, das tatsächlich mit den Plänen für eine verantwortungsvolle Skalierung zusammenhängt, d.h. die US-Regierung hat über die AISI im Grunde Modelle für nationale Sicherheitsrisiken getestet, wie z.B. biologische und nukleare Risiken. Wissen Sie, das Institut ist wahrscheinlich falsch benannt. Sie nennen es das Sicherheitsinstitut. Es klingt nach Vertrauen und Sicherheit. Aber in Wirklichkeit geht es darum, die Risiken für die nationale Sicherheit zu messen. Und wir haben keine Meinung darüber, wo genau das gemacht wird oder wie es genannt wird, aber ich denke, eine Funktion, die diese Messung durchführt, scheint sehr wichtig zu sein.
Es ist auch wichtig, um die Fähigkeiten unserer Gegner zu messen.
Sie können auch die Modelle von DeepSeek messen, um zu sehen, welche Gefahren sie darstellen könnten, insbesondere wenn diese Modelle in den USA verwendet werden. Wozu sind sie in der Lage? Was könnten sie tun, das gefährlich ist? Das ist also Nummer zwei.

Nummer drei, auf der Risikoseite, ist etwas, worüber wir noch nicht gesprochen haben, nämlich dass ich über die Industriespionage von Unternehmen in den USA besorgt bin, Unternehmen wie Anthropic.
Wissen Sie, China ist bekannt für gross angelegte Industriespionage.
Wir machen verschiedene Dinge. In unserem Plan für verantwortungsvolle Skalierung gibt es Dinge wie immer bessere Sicherheitsmassnahmen.
Aber wissen Sie, bei vielen dieser algorithmischen Geheimnisse gibt es **100-Millionen-Dollar-Geheimnisse, die aus ein paar Codezeilen bestehen**.

Und wissen Sie, ich bin mir sicher, dass es Leute gibt, die versuchen, sie zu stehlen, und sie könnten Erfolg haben. Daher ist es sehr wichtig, dass die US-Regierung unsere Unternehmen vor diesem Risiko schützt. Das sind also die drei auf der Sicherheitsseite.

Was die Chancen anbelangt, so ist die – ich denke, **die drei wichtigsten** – **das Potenzial der Technologie**.
In der Anwendungsschicht, in Dingen wie dem Gesundheitswesen.
Ich denke, wir haben, wie gesagt, eine aussergewöhnliche Chance, schwere Krankheiten zu heilen, grosse komplexe Krankheiten, die uns seit Hunderten oder Tausenden von Jahren begleiten und gegen die wir noch nichts tun konnten. Ich denke, das wird auf die eine oder andere Weise passieren, aber die Regulierungspolitik könnte sich wirklich darauf auswirken, ob es fünf Jahre dauert, bis KI uns hilft, all diese Heilmittel herzustellen und an die Welt zu verteilen, oder dauert es dreissig Jahre? Und das ist ein grosser Unterschied für Menschen, die an diesen Krankheiten leiden.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Wir sind der Ansicht, dass die heutige Politik in Bezug auf das Gesundheitswesen, die FDA-Zulassung von Medikamenten, möglicherweise nicht für den schnellen Fortschritt geeignet ist – für den schnellen Fortschritt, den wir sehen werden. Und vielleicht möchten wir einige Hindernisse aus dem Weg räumen.

Die zweite ist die Energieversorgung.
Wenn wir China in dieser Technologie und anderen autoritären Gegnern einen Schritt voraus sein wollen, **müssen wir Rechenzentren bauen. Und es ist besser, wenn wir diese Rechenzentren in den USA oder ihren Verbündeten bauen, als wenn wir sie in Ländern errichten, in denen die Loyalitäten gespalten sind, wo sie buchstäblich mit einem Rechenzentrum fliehen und sagen könnten: "Oh, tut mir leid, wir sind jetzt auf der Seite Chinas.**
Und so wurde einiges davon in den letzten Tagen der Biden-Administration getan. Und ich denke, es ist eine überparteiliche Sache. Ich denke, dass der Trump-Admin, wissen Sie, das ist ein Bereich, in dem man sich einig ist.
Es besteht ein Interesse daran, viel mehr Energie bereitzustellen.

Wir brauchen wahrscheinlich in der gesamten Branche bis 2027 vielleicht fünfzig Gigawatt zusätzliche Energie, um eine KI mit all den Eigenschaften zu versorgen, über die wir gesprochen haben.
Fünfzig Gigawatt, für diejenigen, die es nicht wissen, ist ungefähr die Menge an Energie, die im Jahr 2024 insgesamt in das US-Stromnetz eingespeist wurde. Bis zu diesem Jahr brauchen wir also so viel wie – wissen Sie, halb so viel, wie in den nächsten zwei Jahren hinzukommen wird. Es wird also wirklich viel brauchen.

Und dann ist das Letzte die wirtschaftliche Seite der Dinge.
Wissen Sie, wie wir bereits erwähnt haben, denke ich, **dass die Sorgen auf der wirtschaftlichen Seite genauso existenziell sind wie die Sorgen auf der Seite der nationalen Sicherheit.**

Wissen Sie, kurzfristig werden wir die Disruption bewältigen müssen, auch wenn der Kuchen viel grösser wird. Wissen Sie, auf lange Sicht werden wir, wie ich bereits sagte, über eine Welt nachdenken müssen, in der KI – und ich möchte darüber nicht lügen.

Ich glaube wirklich, dass die KI in fast allen Dingen besser sein wird als fast alle Menschen.
Wir müssen so schnell wie möglich mit dieser Welt rechnen.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Im Moment müssen wir das meiner Meinung nach einfach tun – das Beste, was wir tun können, ist, zu messen, um zu verstehen, was vor sich geht.

Wir haben dieses Ding namens **Anthropic Economic Index veröffentlicht**, das auf eine **datenschutzfreundliche Art und Weise unsere Nutzung durchschaut und zusammenfasst, um zu verstehen,** in welchen Bereichen die Leute es verwenden.
Ist es augmentativ? Ersetzt es? Aber auf lange Sicht werden wir wirklich – wissen Sie, **das wird Fragen zur Steuerpolitik und zur Verteilung des Reichtums mit sich bringen, nicht wahr?**
Es gibt diese Art von verlockender Welt, in der man, wenn der Kuchen genug wächst, die Ressourcen haben könnte, um viel dagegen zu tun.
Nehmen wir mal an – und das wird für dieses Publikum verrückt klingen –, aber nehmen wir an, KI bewirkt ein Wirtschaftswachstum von 10 Prozent pro Jahr. Dann wächst die Steuerbasis plötzlich so stark, dass man das Defizit ausgleichen und vielleicht all das übrig haben kann – all das bleibt übrig, um die wahrscheinlich enorme Disruption zu bewältigen, die von der Technologie ausgeht.
Das wird sich also nach einer verrückten Stadt anhören, aber ich lade Sie einfach ein, das Hypothetische in Betracht zu ziehen und jetzt über die Möglichkeit solcher verrückten Dinge nachzudenken.

FROMAN: Eine verrückte Stadt. **Sie haben es hier zuerst gehört**. OK, lassen Sie es uns für Fragen öffnen. Ja, genau hier vorne.

F: Danke, Dario. Das war ein wirklich faszinierendes Gespräch. Soll ich stehen?

FROMAN: Du stehst einfach da.

F: In Ordnung. Ich werde kandidieren. Mach meine Schritte rein.

FROMAN: Und sagen Sie einfach, wer Sie sind.

F: Ich bin Adem Bunkeddeko. Ich hatte also – und ich habe es genossen, Machines zu lesen – Ihren Essay letztes Jahr, und dann hörte ich Sie über Hard Fork, über die Times, aber auch über dieses. Und die Frage, die ich an Sie habe, ist, wie Sie die politischen und wirtschaftlichen Implikationen skizzieren.

Aber ich bin neugierig darauf, ein Gefühl dafür zu bekommen, wie Sie – wie haben Sie über die sozialen und moralischen Überlegungen nachgedacht, die tatsächlich kommen werden? Vor allem, weil ich denke, dass der Grossteil der Öffentlichkeit einige der Chatbots sieht, etwas davon sieht und sagt, oh, es ist eine verbesserte Google-Suche, aber nicht wirklich über die nachgelagerten Auswirkungen der Disruption auf dem Arbeitsmarkt und dergleichen nachdenkt. Und so bin ich neugierig darauf, ein Gefühl dafür zu bekommen, wie Sie darüber nachdenken, wenn Sie im Spannungsfeld stehen, wenn Sie ein Unternehmen aufbauen, das versucht, ein kommerzielles Produkt zu entwickeln.

AMODEI: Ja. Also, zunächst einmal, ich meine, wissen Sie, ich denke, dass dieses Zeug super wichtig ist. Und vielleicht am meisten – das,

was mich im Moment am meisten beunruhigt, ist das mangelnde Bewusstsein für das Ausmass dessen, was die Technologie wahrscheinlich mit sich bringen wird.

Ich meine, ich könnte mich einfach irren. Ich sage einen Haufen verrückter Sachen. Die Antwort könnte einfach sein, dass die breite Öffentlichkeit Recht hat und ich falsch liege. Ich bin high von meinem eigenen Vorrat.
Ich erkenne an, dass das möglich ist. Aber sagen wir, das ist nicht der Fall.
Was ich sehe, ist, dass es diese konzentrischen Kreise von Menschen gibt, die erkennen, wie gross die Technologie sein könnte.

Es gibt wahrscheinlich ein paar Millionen Menschen – sehr konzentriert im Silicon Valley, aber ein paar Leute hoch in der politischen Welt –, die auch diese Überzeugungen haben.

Auch hier wissen wir noch nicht, ob wir/sie Recht oder Unrecht haben.

Aber wenn wir Recht haben, hält die gesamte Bevölkerung dieses Zeug wieder für Chatbots.

Wenn wir sagen, dass dies gefährlich ist, wenn wir sagen, dass dies alle menschliche Arbeit ersetzen könnte, klingt das verrückt, denn das, was sie sehen, ist etwas, das in einigen Fällen ziemlich frivol erscheint.
Aber sie wissen nicht, was auf sie zukommt. Und ich denke, das ist es, was mich nachts viel wachhält, und das ist der Grund, warum ich versuche, die Botschaft an mehr Menschen weiterzugeben.

Ich denke also, dass Bewusstsein der erste Schritt ist. Ich denke, diese Fragen rund um die menschliche Arbeit und die menschliche Arbeit in einer Welt, in der es technologisch möglich ist, die Auswirkungen des menschlichen Geistes zu replizieren, sind meiner Meinung nach sehr tiefgreifende Fragen.

Ich habe nicht das Gefühl, dass ich die Antwort darauf habe.
Ich habe das Gefühl, wie Sie gesagt haben, das sind – das sind moralische Fragen, fast – wissen Sie, fast, Fragen über den Zweck, man kann sogar sagen spirituelle Fragen, Rechts? Und so werden wir alle gemeinsam diese Fragen beantworten müssen.

Ich meine, ich gebe Ihnen eine Art Embryo einer Antwort, die ich habe, nämlich, dass es irgendwie Aspekte gibt, die tief in unserer Psychologie verankert sind, aber es gibt Aspekte, die kulturell sind.

Weisst du, es gibt viele Dinge, die gut funktionieren.
Es hat eine moderne partizipative Wirtschaft geschaffen. Aber die Technologie, wie so oft, kann diese Illusion irgendwie blosslegen.
Es kann ein anderer Moment sein, wie der Moment, in dem wir erkennen, dass sich die Erde um die Sonne dreht, anstatt dass sich die Sonne um die Erde dreht. Oder, wissen Sie, es gibt viele, viele Sonnensysteme.
Oder organisches Material besteht nicht aus anderen Molekülen als anorganisches Material. Vielleicht haben wir also einen dieser Momente.
Und es könnte eine Abrechnung geben.
Und wieder einmal ist meine Antwort, dass ich beeindruckt bin, wie sinnvoll Aktivitäten sein können, selbst wenn sie keinen wirtschaftlichen Wert generieren. Ich bin beeindruckt davon, wie sehr ich Dinge geniessen kann, in denen ich nicht die Beste der Welt bin. Wenn die Anforderung ist, dass du der Beste der Welt in etwas sein musst, damit es für dich irgendwie spirituell bedeutungsvoll ist, habe ich das Gefühl, dass du eine falsche Abzweigung genommen hast. Ich habe das Gefühl, dass in dieser Annahme etwas falsch ist. Und ich sage das als jemand, der viel Zeit damit verbringt, der Beste der Welt zu sein, wissen Sie, in etwas, das ich für wirklich wichtig halte. Aber irgendwie sind wir – unsere Quelle des Sinns muss etwas anderes sein als das.

FROMAN: Ja, Cam.

F: Danke. Cam Kerry an der Brookings Institution.
Eines der Dinge, die mir im britischen KI-Sicherheitsbericht aufgefallen sind, ist die Möglichkeit, **dass im Jahr 2030 oder so die Daten ausgehen könnten**.

Wie skalieren Sie dann? Wie macht man die Modelle intelligenter?
Und was sind die Grenzen dieser Daten? Ich meine, es gibt eine enorme Menge an Text, Videoinformationen, die digitalisiert sind, eine enorme Menge, die sich in unseren Köpfen und im Universum befindet, die es nicht ist.
Wie gehen Sie damit um?

AMODEI: Ja. Also ein paar Antworten darauf. Eine davon ist, dass es in den letzten sechs Monaten einige Innovationen gegeben hat, die eigentlich nicht von uns entwickelt wurden.
Wissen Sie, die erste kam eigentlich von OpenAI, aber andere, die wir gemacht haben, machen es überflüssig, so viele Daten zu benötigen, wie wir vorher benötigt haben. Das sind die sogenannten Denkmodelle, bei denen sie im Grunde genommen – sie haben Gedanken. Sie beginnen, die Antworten auf komplexe Fragen zu durchdenken. **Und dann trainieren sie ihre eigenen Gedanken**. Man kann darüber nachdenken, wie Menschen das machen, wobei ich manchmal Dinge lernen kann, indem ich – weisst du, ich mache einen Plan in meinem Kopf und dann denke ich noch einmal darüber nach und sage, oh, weisst du, bei genauerem Nachdenken macht das nicht wirklich viel Sinn. Also, was denkst du, oder? Und dann lernt man irgendwie etwas daraus. Natürlich muss man auch in der Welt handeln.
Du musst auch in der realen Welt handeln. Aber KIs haben diese Art der Kognition bis vor kurzem überhaupt nicht genutzt.
Bisher wird dies hauptsächlich auf Aufgaben wie Mathematik und Computer-programmierung angewendet. Aber ich bin der Meinung, ohne zu konkret zu werden, dass es nicht allzu schwierig sein wird, diese Art des Denkens auf ein viel breiteres Spektrum von Aufgaben auszudehnen.

Der zweite Punkt ist, selbst wenn uns im Jahr 2030 die Daten ausgehen, wenn die Exponentialdynamik auch nur zwei oder drei Jahre anhält, könnten wir an einen Punkt gelangen, an dem wir bereits auf der Ebene des Genies sind.
Und Sie wissen, dass das für viele dieser Änderungen ausreichen kann.
Und wir können vielleicht auch die Models fragen: Hey, wir haben dieses Problem. Humanwissenschaftler waren nicht in der Lage, das Problem zu lösen. Können Sie uns helfen, dieses Problem zu lösen? Ich gebe immer noch eine geringe Wahrscheinlichkeit an, dass, aus welchen Gründen auch immer, diese beiden Dinge nicht funktionieren werden oder nicht so sind, wie sie erscheinen, und Daten könnten eines der plausiblen Dinge sein, die uns blockieren könnten. Ich dachte es vor ein oder zwei Jahren für einen sehr plausiblen Blocker.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Vor ein oder zwei Jahren dachte ich, wenn irgendetwas die Show stoppen würde, wäre das unter den ersten drei der Liste der Dinge, die es tun würden. Aber ich denke, dass meine potenzielle Skepsis hier nicht vollständig widerlegt wurde, aber ich denke, dass sie einigermassen gut widerlegt wurde.

FROMAN: **Was sind die drei wichtigsten Dinge, die die Serie stoppen könnten?**

AMODEI: Also, an diesem Punkt denke ich, **dass das Wichtigste, was das stoppen könnte, eine Unterbrechung der Versorgung mit GPUs wäre.** Wenn es zum Beispiel **in dem kleinen, umstrittenen Gebiet, in dem die gesamte GPU hergestellt wird, zu einem militärischen Konflikt käme, wäre das sicherlich der Fall.**
Ich denke, eine andere Sache wäre, **wenn es eine ausreichend grosse Störung am Aktienmarkt gibt, die die Kapitalisierung dieser Unternehmen durcheinanderbringt.**
Im Grunde genommen **eine Art – eine Art Glaube, dass sich die Technologie nicht weiterentwickeln wird, und das schafft eine sich selbst erfüllende Prophezeiung, bei der es nicht genug Kapitalisierung gibt**.

Und drittens würde ich sagen, wenn ich oder wir, das Feld, irgendwie falsch liegen, was die Vielversprechendkeit dieses neuen Paradigmas des Lernens aus den eigenen Daten angeht. Wenn es irgendwie nicht so weit gefasst ist, wie es scheint, oder einfach mehr dahintersteckt, es richtig zu machen, dass unserer Meinung nach einige Erkenntnisse fehlen.

FROMAN: Wir kommen zu einer Online-Frage.

OPERATOR: Wir nehmen die nächste Frage von Esther Dyson.

AMODEI: Ich erkenne diesen Namen.

OPERATOR: Frau Dyson, bitte heben Sie die Stummschaltung Ihrer Leitung auf.

F: Vielen Dank. Entschuldigungen. Esther Dyson, die ein Buch mit dem Titel "Term Limits" über Amtszeitbegrenzungen für Menschen und KIs und so weiter schreibt. **Ich habe eine Frage zu dieser ganzen Sache mit dem existenziellen Risiko.** Es scheint mir, **dass die grösseren Risiken, ehrlich gesagt, von den Menschen ausgehen, die noch unerklärlicher sind als KIs, aber von den Menschen und ihren Geschäftsmodellen, die KIs nutzen.**

Und dann gibt es noch das berühmte Büroklammerproblem, bei dem man die KI bittet, Büroklammern zu erstellen, und sie tut dies unter Ausschluss von allem anderen. Und das ist ein wenig metaphorisch, aber die Welt scheint verrückt nach Rechenzentren zu werden. **Und es zieht wirklich Ressourcen von allem anderen ab, um Rechenzentren, KI, Datenpools und was auch immer zu finanzieren. Und so schafft KI in gewisser Weise eine Fitnessfunktion für die Gesellschaft, die, wie ich denke, dem Wert des Menschen schadet, der nicht nur seine intellektuelle Kapazität ist. Das ist das Ende der Frage. Vielen Dank**.

AMODEI: Also, wissen Sie, ich würde sagen, so wie es viele verschiedene Vorteile von KI gibt – und jedes Mal, wenn wir eine neue KI produzieren – jedes Mal, wenn wir ein neues KI-Modell produzieren, hat sie, wissen Sie, eine lange Liste von zehn Vorteilen, die wir erwartet haben, und dann noch eine Reihe mehr, die wir nicht erwartet haben.
Jedes Mal, wenn wir ein neues Modell veröffentlichen, gibt es neue Anwendungsfälle und die Kunden sagen: Ich habe nicht einmal daran gedacht, das mit einem KI-System zu tun.
Es ist leider auch so, dass es – wissen Sie, wir sollten nicht sagen, **dass dieses Risiko eine Ablenkung von diesem Risiko ist**.
Es ist nur leider so, dass es viele verschiedene Risiken für die KI-Systeme gibt. Und wenn wir das durchstehen wollen, **müssen wir uns irgendwie mit allen auseinandersetzen**.
Ich denke also, dass es ein grosses Risiko ist, dass der Mensch die KI-Systeme missbraucht.
Ich denke, es ist ein grosses Risiko, dass wir bei den KI-Systemen selbst Schwierigkeiten haben könnten, sie zu kontrollieren.
Um die Analogie eines Landes der Genies in einem Rechenzentrum zu verwenden: Wir setzen ein Land mit zehn Millionen Genies in, Sie wissen schon, Antarktis oder so etwas herunter. Wir werden mehrere Fragen darüber haben, was das mit der Menschheit machen wird. Wissen Sie, wir werden fragen, nun, wer – wissen Sie, existiert etwas – besitzt irgendein existierendes Land sie? Ist es – tut es ihren Befehl? Und was bewirkt das?
Wissen Sie, sind die Vorteile – wissen Sie, ist das Ergebnis davon vorteilhaft?
Wir werden sagen, wissen Sie, gibt es Individuen, die es missbrauchen könnten? Und wir werden sagen, was sind die Absichten dieses Landes der Genies selbst? Und um dann zu der Frage zu kommen, die Sie am Ende gestellt haben: Gibt es mehr verteilte gesellschaftliche Dinge? Ich glaube auf jeden Fall, dass, wenn immer mehr Teile der Welt – immer mehr unserer Energie in KI-Systeme investiert wird, das grossartig sein wird. Sie werden die Dinge wirklich effizient erledigen.

Aber könnte das auch einige unserer bestehenden Umweltprobleme verschlimmern? Ich denke, das ist ein echtes Risiko.
Und dann kann man sagen, na ja, werden die KIs uns besser helfen, unsere Umweltprobleme zu lösen? Wir verbrauchen also einen Haufen Energie, und dann die KI – die KI-Systeme, wissen Sie, es stellt sich irgendwie – weisst du, es stellt sich heraus – **wir sind am Ende besser, als wir angefangen haben, wenn wir in der Lage sind, es zu lösen**.
Ich bin also optimistisch, dass das der Fall sein wird, aber das ist ein weiteres Risiko. Es müssen eine Reihe von Dingen wahr sein, damit es sich so herausstellt. Also, wissen Sie, ich glaube einfach, dass wir uns in einer Zeit grosser Veränderungen befinden. Und deshalb, wissen Sie, müssen wir ausserordentlich kluge Entscheidungen treffen, um das durchzustehen.
Ich meine, wissen Sie, ich erkenne den Namen der Person, die die Frage stellt. Und ich könnte – ich könnte das falsch verstehen, aber ich glaube, es war dein – ich glaube, es war dein Vater, der sagte – ich habe mir ein Video von ihm angehört, weil ich – ich war Physiker. Und ich hörte mir ein Video von ihm an, in dem er sagte, wir haben all diese Probleme heute und wir – wissen Sie, es scheint, als könnten wir sie nicht lösen. Aber wissen Sie, ich erinnere mich – ich erinnere mich, dass es zu meiner Zeit wirklich so aussah, als hätten wir all diese schweren, ernsten Probleme, wissen Sie, wenn wir nur an den Zweiten Weltkrieg oder den Kalten Krieg oder die nukleare Vernichtung dachten.
Und irgendwie haben wir es geschafft.
Es bedeutet also nicht, dass wir es wieder tun werden, aber. (Gelächter.)

FROMAN: Ja. Die Frau hinten, dort.

F: Hallo. Mein Name ist Carmem Domingues. Ich bin ein KI-Spezialist mit einem Hintergrund in der Entwicklung, Implementierung und in letzter Zeit mit einem stärkeren Fokus auf die politische Seite.
Ich höre laut und deutlich den Mangel an Bewusstsein im Allgemeinen darüber, was KI ist und was nicht und was sie kann und was nicht.
Aber das überspringe ich. Ich mache auch Wissenschaftskommunikation zu diesem Thema. Aber meine Frage heute ist, dass Sie vor ein paar Monaten Kyle Fish als KI-Wohlfahrtsforscher engagiert haben, um sich mit dem Empfindungsvermögen oder dem Fehlen eines solchen von zukünftigen KI-Modellen zu befassen und ob sie in Zukunft moralische Berücksichtigung und Schutz verdienen könnten. Wenn Sie ein wenig darüber sprechen könnten, die Gründe dafür und **ob Sie ein gleichwertiges Forschungsteam für das Wohlergehen des Menschen haben**. Danke.

AMODEI: Ja. Das ist also – das ist wieder eines dieser Themen, bei denen ich völlig verrückt klingen werde. Ich bin also der Meinung, dass, wenn wir diese Systeme bauen, und Sie wissen schon, sie sich in vielen Details von der Art und Weise unterscheiden, wie das menschliche Gehirn aufgebaut ist, aber die Anzahl der Neuronen, die Anzahl der Verbindungen, ist auffallend ähnlich. **Einige der Konzepte sind sich auffallend ähnlich. Ich habe eine – ich habe eine funktionalistische Auffassung von, wissen Sie, moralischem Wohlergehen der Natur der Erfahrung, vielleicht sogar des Bewusstseins**.
Und so denke ich, dass wir zumindest die Frage in Betracht ziehen sollten, ob wenn wir diese Systeme aufbauen und sie alle möglichen Dinge tun wie Menschen und sie alle möglichen Dinge tun wie Menschen und viele der gleichen kognitiven Fähigkeiten zu haben scheinen, wenn sie wie eine Ente quakt und wie eine Ente läuft, dann ist es vielleicht eine Ente.
Und wir sollten wirklich darüber nachdenken, ob diese Dinge echte Erfahrungen machen, die in irgendeiner Weise bedeutungsvoll sind.
Wenn wir Millionen von ihnen einsetzen und nicht über die Erfahrungen nachdenken, die sie haben, und sie haben vielleicht keine. Das ist eine sehr schwer zu beantwortende Frage. Das ist etwas, worüber wir sehr ernsthaft nachdenken sollten. Und das ist nicht nur eine philosophische Frage.
Ich war überrascht zu erfahren, dass es überraschend praktische Dinge gibt, die man tun kann. Also, wissen Sie, etwas, worüber wir nachdenken, wenn wir mit der Bereitstellung beginnen – wenn wir unsere Modelle in ihren Bereitstellungsumgebungen bereitstellen, indem wir dem Modell einfach eine Schaltfläche geben, die sagt: "Ich habe diesen Job gekündigt", die das Modell drücken kann, richtig? Es ist nur eine Art sehr grundlegender Präferenzrahmen, bei dem man sagt, wenn man davon ausgeht, dass das Modell Erfahrung hat und den Job genug hasst, was ihm die Möglichkeit gibt, den Knopf zu drücken: "Ich habe diesen Job gekündigt." Wenn Sie feststellen, dass die Models diesen Knopf oft für Dinge drücken, die wirklich unangenehm sind, sollten Sie vielleicht etwas bezahlen – das bedeutet nicht, dass Sie überzeugt sind, aber vielleicht sollten Sie dem etwas Aufmerksamkeit schenken. Klingt verrückt, ich weiss. Es ist wahrscheinlich das Verrückteste, was ich bisher gesagt habe.

FROMAN: Ganz hinten dort. Trooper, ja.

F: Hallo. Soldat Sanders. Sie haben über die Begeisterung für KI und Medizin, Biologie, Chemie und so weiter gesprochen. **Wenn Sie was sagen könnten – gibt es irgendeine Aufregung in den Sozialwissenschaften**?

Der grösste Teil der Gesundheitsversorgung findet ausserhalb der Pillendose und des Untersuchungsraums statt. Die öffentliche Gesundheit umfasst eine Reihe weiterer Bereiche. Können Sie dazu etwas sagen?

AMODEI: Ja. **Ich meine, wenn ich an Epidemiologie denke, wissen Sie, als ich an der Universität war, gab es ein Projekt der Gates Foundation, bei dem es darum ging, mathematische und computergestützte Methoden rund um die Epidemiologie zu verwenden.**
Ich glaube, sie planten, es zu nutzen, um Malaria, Polio und andere Bereiche auszurotten. Die Menge an Daten, die wir erhalten, und die Fähigkeit, alle Teile zusammenzufügen und zu verstehen, was in einer Epidemie vor sich geht, könnte sicherlich enorm von KI profitieren. Der Ablauf der klinischen Studie. So etwas haben wir schon gesehen. Das ist also etwas, was Anthropic mit Novo Nordisk gemacht hat, dem Hersteller von Ozempic und anderen Medikamenten. Am Ende einer klinischen Studie müssen Sie einen klinischen Studienbericht verfassen.
Wissen Sie, es fasst unerwünschte Vorfälle zusammen, führt alle statistischen Analysen durch, um sie der FDA oder anderen Aufsichtsbehörden vorzulegen, um zu entscheiden, ob das Medikament zugelassen werden soll. In der Regel dauert dies etwa zehn Wochen. Sie haben begonnen, unser Modell dafür zu verwenden. Und das Modell braucht etwa zehn Minuten, um den klinischen Studienbericht zu schreiben, und der Mensch braucht etwa drei Tage, um ihn zu überprüfen. Und die Qualität, zumindest wie wir sie in frühen Studien gesehen haben – die nicht alles bestimmt – wurde als vergleichbar mit dem angesehen, was der Mensch mit dem zehnwöchigen Prozess zu tun imstande ist. **Wir müssen also klinische Studien durchführen**.
Es gibt eine Menge sozialwissenschaftlicher Probleme in diesem Zusammenhang. Es gibt eine Menge regulatorischer Probleme. Ich schreibe in dem Essay ein wenig darüber, dass ich denke, dass diese Dinge das Tempo des Fortschritts begrenzen werden. **Aber selbst bei Dingen wie klinischen Studien denke ich, dass die KI-Systeme in der Lage sein werden, diese Fragen zwar nicht zu lösen, aber zumindest radikal zu vereinfachen.**

FROMAN: Ja, genau hier. Da kommt ein Mikrofon.

F: Ich bin Louise Shelley. Ich bin ein Experte auf dem. Illegaler Handel, von der George Mason University. **Nächste Woche findet ein Weltgipfel der OECD zum illegalen Handel statt.** Aber das, worüber Sie gesprochen haben, ist nicht das, was ich zu diesem Problem des Teileschmuggels erwartet hatte.

Und es ist auf niemandes Radarschirm. Was passiert, wenn du darüber sprichst? Weil es nicht die Gemeinschaft erreicht, die diesen illegalen Handel schützen muss.

AMODEI: Ich habe den letzten Teil der Frage nicht gehört.

FROMAN: Wie kommt es also, dass diese Themen nicht auf der Tagesordnung **der Leute stehen, die sich Sorgen um den illegalen Handel machen?**

AMODEI: Ja. Weisst du, ja. Ich denke, meine Antwort darauf ist, wissen Sie, es sollte auf dem Radar dieser Leute sein.
Weisst du, noch einmal, ich habe hier eine Weltanschauung, die nicht jeder teilt. Und ich kann Recht haben oder ich kann mich irren. Aber alles, was ich sagen kann, ist, wenn diese Weltanschauung richtig ist, dann sollten wir uns viel mehr Sorgen um den Schmuggel dieser GPUs machen als um den Schmuggel von Waffen oder sogar Drohnen oder Fentanyl oder was auch immer.
Ja, aber – wissen Sie, wenn Sie fünf Millionen davon nach China schmuggeln würden – und um es klar zu sagen, das sind etwa 20 Milliarden Dollar Wert oder so etwas in der Art – dann würde das das nationale Sicherheitsgleichgewicht der Welt drastisch verändern. Ich denke, es ist das – ich denke, es ist das Wichtigste.
Also, wissen Sie, das ist – das ist das Dilemma: Bin ich einfach verrückt oder **hat die Welt hier ein grosses, grosses Bewusstseinsproblem?**
Und wenn die Welt hier ein grosses, grosses Bewusstseinsproblem hat, dann ist eine nachgelagerte Konsequenz davon, dass wir uns auf all diese anderen Dinge konzentrieren. Und wissen Sie, denn wenn man von illegalem Handel spricht, gibt es bestimmte Dinge, auf die sich die Leute schon lange konzentrieren, das ist eine neue Sache. Aber das bedeutet nicht, dass es nicht das Wichtigste ist.

FROMAN: Ah, so viele gute Fragen, da bin ich mir sicher. Dieser Herr hier.

F: Vielen Dank. Alan Raul. **Praktizierender Anwalt, Dozent an der Harvard Law School und zukünftige nutzlose Person. (Gelächter.)**

AMODEI: **Das sind wir alle auch.**

F: **Ich möchte also an Ihre verschiedenen Kommentare zur nationalen Sicherheit anknüpfen.** Sie haben das Artificial Intelligence Security Institute und dessen Tests erwähnt.

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Die Biden-Durchführungsverordnung zu KI sah eine obligatorische Berichterstattung über den Erwerb oder die Entwicklung von superfähigen und 26 FLOP-Stiftungsmodellen mit doppeltem Verwendungszweck vor.
Aber meine Frage ist, wie engagieren Sie sich? Wie gehen Anthropic, die KI-Community, die Entwickler dieser superfähigen Modelle, wie gehen sie mit der nationalen Sicherheitsgemeinschaft der USA um, wie gehen sie mit den Geheimdiensten um? Und was bedeutet das für die Entwicklung von KI in der Praxis? Und wenn du mir sagst, dass du mich umbringen müsstest, muss ich es nicht so schlimm wissen. (Gelächter.)

AMODEI: Ja. Ich denke also, dass es hier ein paar Dinge gibt.
Eines ist typisch Anthropic, obwohl die anderen Unternehmen begonnen haben, ähnliche Dinge zu tun, haben wir jedes Mal, wenn wir ein neues Modell entwickeln, ein Team innerhalb von Anthropic namens **Frontier Red Team**. **Einiges davon findet in Tests mit den AI Safety and Security Institutes statt, aber wir arbeiten zusammen.** Wir entwickeln einige Sachen. Sie entwickeln einige Sachen. Aber der allgemeine Ablauf war, wenn **wir die Modelle auf Dinge wie biologisches Risiko oder Cyberrisiko testen, oder, Sie wissen schon, chemisches oder radiologisches Risiko, gehen wir in der Regel zu Leuten in der nationalen Sicherheitsgemeinschaft und sagen: Hey, das ist der Punkt, an dem sich die Modelle in Bezug auf diese speziellen Fähigkeiten befinden. Wissen Sie, Sie sollten darüber Bescheid wissen, denn Sie sind diejenigen, die dafür verantwortlich sind, die schlechten Akteure zu erkennen, die dies mit den Modellen tun würden. Du weisst jetzt, wozu sie fähig sind. Vielleicht haben Sie also ein Gefühl dafür, was die Modelle leisten können, was additiv oder ergänzend zu ihren aktuellen Fähigkeiten ist, oder? Der Teil, den wir übersehen, ist, dass wir keine Experten für Terrorismusbekämpfung sind.**

Wir sind keine Experten für alle Bösewichte auf der Welt und darüber, was ihre Fähigkeiten sind und was grosse Sprachmodelle dem Bild hinzufügen würden. Und so haben wir einen sehr produktiven Dialog mit ihnen über diese Themen geführt.

Das andere Thema, über das wir mit ihnen gesprochen haben, ist die Sicherheit der Unternehmen selbst. Wissen Sie, das war eines der Dinge in unserer Art von OSTP-Einreichung, wissen Sie, das formeller zu machen, das zu etwas zu machen, was die US-Regierung als Selbstverständlichkeit tut. Aber wissen Sie, wenn wir befürchten, dass wir digital angegriffen werden oder – oder durch eine Art von Insider-Bedrohung, wissen Sie, mit menschlichen

Mitteln, dann werden wir oft mit der nationalen Sicherheitsgemeinschaft darüber sprechen.

Und ich denke, die dritte Art der Interaktion dreht sich um die Auswirkungen der Modelle auf die nationale Sicherheit, richtig? Wir waren – wissen Sie, diese Dinge, die ich jetzt öffentlich sage, ich war – wissen Sie, ich habe sie in irgendeiner Form schon eine ganze Weile zu einigen Leuten gesagt.

Und dann denke ich, **dass der vierte Punkt ist, dass es eine Möglichkeit gibt, die Modelle anzuwenden, um unsere nationale Sicherheit zu verbessern.**
Das ist etwas, das ich und Anthropic unterstützt haben, obwohl wir sicherstellen wollen, dass es die richtigen Leitplanken gibt, oder?
Auf der einen Seite denke ich, dass wir, wenn wir diese Technologien nicht für unsere nationale Sicherheit einsetzen, unseren Gegnern schutzlos ausgeliefert sein werden.

Auf der anderen Seite glaube ich, dass jeder der Meinung ist, dass es Grenzen geben sollte. Wissen Sie, ich glaube nicht, dass es irgendjemanden gibt, der denkt, wir sollten – wir sollten, wissen Sie, KI-Systeme an Atomwaffen anschliessen und sie Atomwaffen abfeuern lassen, ohne dass Menschen auf dem Laufenden sind. Das ist die Handlung von Dr. Strangelove. Ja, das ist buchstäblich die Handlung von Dr. Seltsame Liebe.
Irgendwo dazwischen gibt es also etwas – wie, wissen Sie, da ist etwas Boden. Und wissen Sie, wir arbeiten immer noch daran, das zu definieren.
Es ist eines der Dinge, bei denen wir hoffen, eine Art Vorreiterrolle bei der Definition dessen zu übernehmen, was der angemessene Einsatz von KI für die nationale Sicherheit ist. Aber das ist ein weiterer Bereich, in dem wir mit der nationalen Sicherheitsgemeinschaft interagiert haben.

FROMAN: Ihr Kommentar vor ein paar Minuten, dass Sie versucht haben, die Erfahrung der KI-Modelle zu verstehen, ist mir irgendwie in den Sinn gekommen. Lassen Sie mich also mit einer letzten Frage schliessen.
Was bedeutet es in der Welt, die du dir vorstellst, ein Mensch zu sein?

AMODEI: Ja. Weisst du, ich denke, mein Bild davon – das, was scheint – das, was am meisten scheint – **es gibt vielleicht zwei Dinge, die mir am menschlichsten erscheinen.**

Die Zukunft der KI-Führung in den USA mit CEO von Anthropic Dario Amodei

Das erste, was mir am menschlichsten erscheint, ist, dass wir uns durch unsere Beziehungen zu anderen Menschen kämpfen, durch unsere Verpflichtungen ihnen gegenüber, wie wir sie behandeln müssen, die Schwierigkeiten, die wir in unseren Beziehungen zu anderen Menschen haben und wie wir diese Schwierigkeiten überwinden.

Weisst du, wenn ich an Dinge denke, auf die die Leute stolz sind, und an die grössten Fehler, die sie gemacht haben, dann beziehen sie sich fast immer darauf. Und KI-Systeme können uns vielleicht helfen, das besser zu machen, aber ich denke, **das wird immer eine der wesentlichen Herausforderungen des Menschseins sein**.

Und ich denke, **die zweite Herausforderung ist vielleicht der Ehrgeiz, sehr schwierige Dinge zu tun, die, ich wiederhole es nochmals, letztendlich nicht von der Existenz von KI-Systemen beeinflusst werden, die intelligenter sind als wir und Dinge tun können, die wir nicht tun können. Ich denke wieder daran, dass menschliche Schachmeister immer noch Prominente sind**.

Weisst du, ich kann schwimmen lernen oder Tennis spielen lernen, und die Tatsache, dass ich nicht Weltmeister bin, **negiert nicht die Bedeutung dieser Aktivitäten.** Und wissen Sie, sogar Dinge, die ich über fünfzig Jahre, über hundert Jahre hinweg tun könnte, wissen Sie, **ich möchte, dass diese Dinge ihre Bedeutung behalten und, wissen Sie, die Fähigkeit der Menschen, nach diesen Dingen zu streben, nicht aufzugeben**.
Weisst du, noch einmal, ich denke – ich denke, diese beiden Dinge sind vielleicht das, was ich identifizieren würde.

FROMAN: Bitte schliessen Sie sich mir an und danken Sie Dario Amodei dafür, dass er Zeit mit uns verbracht hat. (Beifall.)

AMODEI: Danke, dass Sie mich eingeladen haben. (ENDE)