Quelle: derStandard.de
Website: https://www.derstandard.de/
Interview: https://www.derstandard.de/story/3000000215426/ki-experte-wir-werden-zur-zweitintelligentesten-spezies
Datum des Interviews: Mai 2024
Interviewpartner: Gary Marcus
Interviewer: derStandard-Redaktion

Das Rennen um superintelligente Maschinen spitzt sich zu. Es geht um Geld und Macht und viel, was schiefgehen kann, erklärt Patrick Swanson im Interview
Geht es nach den Herstellern von Künstlicher Intelligenz, stehen wir am Beginn einer neuen Ära: Schon in wenigen Jahren sollen Maschinen Menschen als intelligenteste Spezies auf dem Planeten ablösen. Ein Szenario, das auch außerhalb der Marketingschmieden des Silicon Valley in nicht allzu ferner Zukunft eintreten könnte, erklärt KI-Experte Patrick Swanson im Interview mit dem STANDARD. Doch wer wird die erste Super-KI kontrollieren? Und, wie bringen wir sie dazu, dass sie nicht zum Untergang der Menschheit führt?

Interviewer*in: Welche Jobs hat KI bereits von uns übernommen?
Interviewee: Wenn du in San Francisco lebst, dann weißt du, dass selbstfahrende Autos nicht die Zukunft sind, sondern etwas, dass es schon gibt. Ich sehe sie jeden Tag auf der Straße, es werden jeden Tag mehr. Das ist ein Bild dafür, wie sich rund um dich herum die Welt ändert und die KI Aufgaben übernimmt. Und es wird sich jeder Job, der an einem Computer ausgeführt wird, in den nächsten Jahren massiv verändern. Aufgaben wie zum Beispiel das Zusammenfassen von langen Dokumenten. Google Maps ist nichts anderes als KI. Das ist maschinelles Lernen dafür, wie wir am schnellsten zum Ziel kommen. Und generative künstliche Intelligenz wird in den nächsten Jahren immer mehr Aufgaben von uns übernehmen. Die Frage ist, ob wir das nur nutzen wollen, um effizienter und kostengünstiger zu werden. Mit massiven Auswirkungen auf den Arbeitsmarkt. Oder, ob wir das für neue Dinge nutzen wollen. Ich finde, es ist nicht sinnvoll zu sagen, wir machen weiter wie bisher, nur immer schlechter und billiger.
Interviewer*in: Vor kurzem hat US-Präsident Donald Trump weitreichende Zölle gegen diverse Länder verhängt oder angedroht. Im Zuge dessen kam ein Gerücht auf, wonach die Bemessungsgrundlage für diese Zölle von ChatGPT geschrieben wurde. Greift KI in den USA tatsächlich schon so weit ein, dass sie das Leben auf so bedeutende Weise mitbestimmt?
Interviewee: In fast jedem Job hier greift KI massiv ein. Die Trump-Zölle sind ja ein lustiges bzw. beängstigendes Beispiel dafür. Das kommt daher, weil jemand eine Analyse gemacht und gesehen hat, was passiert, wenn man KI-Tools darum bittet, eine möglichst simple Formel für Zölle zu berechnen. Und da kam so eine Formel heraus, wie sie jetzt verwendet wurde. Die Vermutung ist, dass Peter Navarro, ein Berater im Weißen Haus, der für die Zölle zuständig ist, früher Texte geschrieben hat darüber, wie man Zölle berechnen könnte. Das könnte in die Trainingsdaten eingeflossen sein. So genau wissen wir es nicht, aber KI wird in der Politik natürlich verwendet werden.
Interviewer*in: Wo liegen aktuell noch die Grenzen von KI?
Interviewee: Es gibt Dinge, die KI heute nicht kann und die sie auf längere Sicht nicht können wird. Journalisten beispielsweise gehen raus, reden mit Menschen und bauen zu anderen Menschen Vertrauen auf. Daraus werden dann Quellen, etwa für investigative Recherchen. Das werden wir von KI noch länger nicht bekommen, weil es natürlich einen emotionalen Kern gibt, den wir als Menschen haben und auch letztlich einen Körper. Und KI hat noch weitere Limitationen. Es gibt Halluzinationen, Vorurteile, unterschiedliche Probleme, die diese Modelle haben.

Programmieren hingegen könnte nächstes Jahr wahrscheinlich zu einem großen Teil automatisiert werden. Das ist jedenfalls das, was die Chefs von diesen großen KI-Firmen sagen, dass dieses und nächstes Jahr beim Programmieren riesige Sprünge passieren werden. Bei Kreativität, Text und Bildern passiert das schleichend nebenbei. Und Kreativität ist etwas, wo ich optimistisch bin, dass wir als Menschen eine positive Entwicklung vor uns haben. Und dann gibt es andere Tätigkeiten wie zum Beispiel jene von Automechanikern, wo wir noch weit davon entfernt sind, diese mit KI ersetzen zu können.
Interviewer*in: Weil die physische Komponente komplexer ist?
Interviewee: Genau. Was wir tun, ist letztlich, einen Körper zu bauen. Wir haben mit großen Sprachmodellen wie ChatGPT das Sprachzentrum entwickelt. Also kann es jetzt Text generieren und Sprache verstehen. Dann haben wir Augen, weil diese Modelle mittlerweile sehen können. Sie können Videos analysieren, sie können Bilder analysieren. Diese Textmodelle können auch Sprache artikulieren, also reden. Und so ist quasi der Kopf fertig. Der physische Teil, der Körper, der Roboter, der ist noch recht weit entfernt. Aber am Ende landet man bei einem Replika des Menschen.
Interviewer*in: Ich habe in den letzten zwei Jahren mithilfe von KI für meine Kinder diverse Geschichten geschrieben und dabei erlebt, wie gravierend sich KI verbessert hat in der Art und Weise, wie man lebendige, spannende Geschichten erzählt. Gleichzeitig frage ich mich, was das für uns Menschen bedeutet. Braucht es uns als Erzeuger von kreativen Werken in Zukunft noch?
Interviewee: Ich glaube schon. Und das ist eine Überzeugung, die geht gegen den Mainstream. Es gibt eine Studie zu KI und Memes, also diese Online-Bilder. Die Frage war: Kann KI Memes erstellen, die lustiger, kreativer sind als jene von Menschen? Die Studie kommt zum Ergebnis, dass im Durchschnitt KI-Memes lustiger, kreativer, "more shareable" sind als die menschlichen Memes. Im Durchschnitt, was unbestritten arg ist. Aber was die Studie auch zeigt: Die Highlights, die allerbesten, die lustigsten Memes, die sind menschlichen Ursprungs oder menschengemacht mit KI-Unterstützung. Und ich glaube, dass das die Welt ist, in die wir gehen. In der Menschen, die nicht so kreativ sind, von KI auf ein durchschnittliches Level gehoben werden. Aber wenn jemand bereits ein toller Kabarettist oder Satiriker ist, wird die menschliche Komponente weiterhin eine große Rolle spielen. Weil es einen Unterschied macht, dass unsere Geschichten menschlich sind. Wir könnten schon jetzt tausende Taylor-Swift-Songs generieren. Die würden uns vielleicht 24 Stunden lang interessieren, und dann aber nicht mehr. Weil das, was wir ja wirklich wollen, sind Taylor Swifts Geschichten und Taylor Swifts Herzschmerz. Wir wollen zu dieser Person eine Verbindung haben. Deswegen zahlen wir so viel für die Konzerttickets, damit wir die Person sehen. Und ich glaube, dass dieser Wert nicht weggehen wird. Ich glaube, dass der sogar wichtiger wird. Denn in einer Welt, in der wir überschwemmt werden von KI-generierten Inhalten und alles hochwertig, aber auch generisch aussieht, in der wird der menschliche Teil für die Kunst essenziell bleiben.
Interviewer*in: Viel diskutiert wird in diesem Zusammenhang die Frage, wer für die Arbeit von KI entlohnt werden soll. Haben wir ein System, das Urheber, die ganz viele von diesen KI-Trainingsdaten liefern, entlohnt?
Interviewee: Nein und das ist ein großes Problem. KI-Modelle werden auf Basis des ganzen Internets trainiert. Die Trainingsdaten werden einfach "gescraped", da wird nicht lange nachgefragt. Das basiert auf einer Auslegung des amerikanischen Urheberrechts. Das amerikanische Urheberrecht hat sehr weite Möglichkeiten für Fair Use, für Transformation von Werken und Remixes. Und jetzt sagen KI-Firmen, das trifft auch auf unser KI-Training zu. Deswegen dürfen wir das machen, das ist legal. Alle Artikel, die jemals geschrieben wurden, sind drin. Bücher

sind drin. Studien sind drin. Bilder und Kunst sind drin – für all das wurden die Urheber nicht entlohnt. Jetzt ist die Frage, ist unsere Gesetzgebung hier auf dem aktuellen Stand? Und wie gehen wir damit um, wenn in Zukunft möglicherweise Leute ihre Jobs verlieren, weil eine KI die Arbeit dieser Menschen so gut machen kann wie die Menschen selbst? Das ist auch eine offene Rechtsfrage. Die New York Times zum Beispiel hat Open AI geklagt, weil es ihrer Ansicht nach eine riesige Urheberrechtsverletzung ist, dass ihre Artikel in ChatGPT sind. Sie wollen dafür bezahlt werden. Wenn die New York Times gewinnt, dann wird es, glaube ich, gigantische Verfahren geben mit gigantischen Entschädigungszahlungen an jeden, der sich anschließt. Wenn dieses Verfahren so ausgeht, dass Open AI im Recht ist, dann ist das ein Präzedenzfall dafür, dass wahrscheinlich niemand entlohnt wird. Das stellt auch viele im Silicon Valley vor ein Dilemma. Denn es sind nicht nur Künstler bedroht, sondern auch Programmierer. Auch deren Code ist in den Trainingsdaten. Und im Silicon Valley ist die Antwort darauf dann sehr oft: bedingungsloses Grundeinkommen. Wir werden mit dieser Technologie so viel Wertschöpfung erzielen, sie wird so mächtig sein, sie wird so wahnsinnig viel Produktivität kreieren, dass wir die Erträge daraus dann irgendwie umverteilen müssen. Vielleicht ist es eh eine gute Idee, aber ich finde, es ist auch eine Abwälzung der Verantwortung.
Interviewer*in: Angenommen, es verlieren künftig sehr viele Menschen ihre Arbeit und schaffen keine kreativen Werke und keine Codes und Texte mehr. Was passiert dann mit den KI-Systemen, die bisher abhängig waren von menschlichen Trainingsdaten?
Interviewee: Das wird jetzt niemanden glücklich machen, aber die menschlichen Daten sind für die KI-Modelle nicht mehr so wichtig. Die ganzen Trainingsdaten sind alle schon drin, und heute kann man synthetische Daten generieren. Man kann also mit KI jetzt 15.000 neue Picassos generieren und diese dann wieder als Trainingsdaten verwenden. In Zukunft wird das Training von KI-Modellen zu einem guten Teil auf KI-generierten Daten basieren. Das trifft nicht auf alle Bereiche zu. Was zum Beispiel den Journalismus auszeichnet, sind Echtzeitdaten. Wir wissen als Erstes, wenn etwas passiert, weil wir möglicherweise eine Quelle im Weißen Haus haben. Diese Echtzeitdaten sind wertvoll, und ich glaube, da liegt auch die Verhandlungsmasse, die der Journalismus gegenüber KI-Firmen hat.
Interviewer*in: Die Hoffnung, die viele Menschen haben, ist, dass wir mit KI Dinge schaffen werden, die uns alleine nicht gelingen. Zum Beispiel der Kampf gegen Krebs, die Weltraumforschung, die Entwicklung neuer Materialien. Welche Mammutaufgaben werden wir mit KI meistern?
Interviewee: Beginnen wir mit Bildung. Wir haben jetzt schon die Möglichkeit, jedem Kind einen KI-Tutor zur Seite zu stellen und alle möglichen Fragen und Hausaufgaben und Interessen von Kindern in Echtzeit personalisiert beantworten zu können. Eine omnipräsente Nachhilfe. Das macht einen riesigen Unterschied. Und es gibt erste Studien von Schulen, die damit experimentieren, die zu wirklich guten Ergebnissen kommen. Ich glaube, wir müssen das gesamte Bildungssystem neu denken. Weil einerseits viele von den alten Aufgaben, wie Essays zu schreiben, durch ChatGPT hinfällig geworden sind. Aber vor allem auch deshalb, weil wir damit unfassbare neue Möglichkeiten haben. Und wir müssen uns überlegen, wie wir KI einbauen, damit KI nicht einfach nur dazu führt, dass man einfach alles die KI machen lässt. Aber der richtige kritische Umgang mit KI-Systemen und das Lernen damit halte ich für eine riesige Chance.Das zweite Feld ist die psychologische Gesundheit. Wir haben ein riesiges Problem, dass viele Menschen sich Psychotherapie nicht leisten können, aber Unterstützung brauchen. Es gibt viele Studien mittlerweile, die zeigen, dass man KI- Chatbots zur Unterstützung von Therapieprozessen verwenden kann. Ich finde, das ist ein sehr lebensbejahender Gedanke, dass man sich helfen lassen kann. Und zwar nicht nur die eine Stunde in

der Woche, die ich mir leisten kann, sondern dass ich vielleicht einen auf Therapie spezialisierten Chatbot hätte, der mir als Gesprächspartner zur Verfügung steht und mir hilft, Dinge einzuordnen und über mich selbst nachzudenken. Das wird Leben retten.
Interviewer*in: Ich habe sofort den Film Her vor mir, in dem sich ein Schriftsteller in eine Künstliche Intelligenz verliebt. Für mich ist der Gedanke, eine Beziehung mit einer KI aufzubauen, gruselig. Denn man ist nicht weniger allein, wenn man mit einer KI spricht. Ein echter sozialer Kontakt ist es nicht.
Interviewee: Völlig richtig. Wenn ich mit einem Chatbot spreche, ist das nicht das Gleiche wie menschlicher Kontakt. Und wir sollten vor allem aus vielen Jahren mit Social Media gelernt haben, dass eine Textnachricht von jemandem nicht gleichbedeutend mit einem menschlichen Kontakt ist. Hier ist aber die Frage, benutzen wir diese Chatbots als Ersatz oder als Zusatz? Wenn es ein Ersatz ist für menschliche Interaktion, haben wir ein Problem. Weil dann vereinsamen wir immer weiter mit unseren Computern. Das kann nicht die Zukunft sein, die wir wollen. Wenn ich mir von einem Chatbot aber dabei helfen lassen kann, rauszugehen in die Welt und Freunde zu finden, wenn er mir in sozialen Situationen, vor denen ich mich fürchte, Tipps gibt, dann glaube ich, dass das hilfreich ist. Wenn die Zukunft ist, dass wir nur noch mit Virtual-Reality-Brillen zu Hause sitzen und mit KI-Avataren sprechen, dann ist es schiefgelaufen.
Interviewer*in: Bleiben wir noch kurz bei positiven Zukunftsaussichten. Wie kann KI die Forschung positiv vorantreiben?
Interviewee: Fangen wir mal mit der Medizin an, da haben wir ganz offensichtliche Felder wie zum Beispiel die Analyse von Röntgenbildern, wo Künstliche Intelligenz wahnsinnig gut ist und schneller und präziser als menschliche Ärzte ist. Genauso in der Diagnostik. Es gab eine Studie kürzlich, in der "Ärzte", "Ärzte mit ChatGPT" und "ChatGPT allein" unterschiedliche Krankheitsbilder diagnostizieren mussten auf Basis einer Symptombeschreibung. Und das Ergebnis war, dass die Ärzte allein 75 Prozent Genauigkeit hatten. Die Ärzte mit ChatGPT hatten 76 Prozent Genauigkeit und ChatGPT allein hatte eine Genauigkeit von 95 Prozent. Das bedeutet, die KI allein war in dem Fall am präzisesten, und es war gefährlich, dass der menschliche Arzt mit dabei war. Das nächste Ziel ist jede Form von Forschung im naturwissenschaftlichen Bereich oder was den Klimawandel betrifft. Was die KI-Firmen aber am meisten interessiert, ist die Frage, wie wir KI dazu bringen, KI-Forschung zu betreiben. Weil wenn wir die KI-Forschung automatisieren können, dann kommen wir in einen Selbstverbesserungszyklus, in dem KI sich selbst weiterentwickelt. Das nennt man den Intelligence-Take-off. Das würde uns schnell zu optimistischen Szenarien führen, in denen wir zum Beispiel Krebs heilen können, und auf der anderen Seite zu sehr bedrohlichen Szenarien.
Interviewer*in: Es gibt ein Videospiel namens Universal Paper Clips. Da geht es darum, eine KI zu entwickeln mit dem Ziel, Heftklammern zu produzieren. Spoiler: Bis die KI am Schluss alle Ressourcen der Welt aufbraucht.
Interviewee: Ja genau, und das klingt nach so einem Science-Fiction-Spiel, aber das ist ein echtes KI-philosophisches Problem, an dem Menschen arbeiten. Was bedeutet das, wenn wir eine KI darauf trainieren, ein Ziel zu erreichen? Wenn wir sagen, wir wollen so viele Heftklammern wie möglich produzieren, und die KI aber nicht den Wert von menschlichen Leben sieht. In extremer Rationalität könnte das so weit gehen, dass die KI sagt: Menschen brauche ich nicht für meine Heft-Klammern, also brauchen wir sie nicht mehr.
Interviewer*in: Wie weit sind wir denn davon entfernt, eine KI zu entwickeln, die in allen Bereichen klüger und besser ist als wir Menschen?
Interviewee: Das geht in zwei Stufen. Das erste ist eine KI, die im Grunde so gut ist wie Menschen, die so intelligent ist wie wir. Das nennt

man Allgemeine Künstliche Intelligenz – Artificial General Intelligence auf Englisch, kurz AGI. Und hier haben wir im letzten halben Jahr ernstzunehmende Fortschritte gemacht. Die KI-Labs selbst haben ihre Prognosen verkürzt. Die KI-Firma Anthropic hat kürzlich ein Paper veröffentlicht zur Beratung der US-Regierung, und in diesem Paper steht drin, dass im Herbst 2026 oder Winter 2027 intelligente KI-Systeme entstehen werden, die äquivalent sind zu Nobelpreisträgern in allen Bereichen. Und die auch das Internet benutzen können, die Computer benutzen wie wir. Zu denen man sagen kann, ich habe eine Rechercheaufgabe, bitte erledige das für mich, und die KI meldet sich dann wieder, wenn sie ihre Aufgabe gelöst hat oder Rückfragen hat. Letztlich wie ein Mitarbeiter oder eine Mitarbeiterin. Das ist die Vision von KI-Firmen. Und die schlagen relativ laut Alarm: Wir sehen bei unseren Modellen, dass wir schnelle Fortschritte machen, und wir müssen das ernst nehmen, weil das wahnsinnig disruptiv ist. Jetzt kann man sagen, das ist Marketing. Die wollen halt ihr Produkt verkaufen. Weil, wenn ich sage, ich habe eine KI, die so mächtig ist, dann sage ich damit auch, bitte investiert in mich. Also das muss man schon skeptisch betrachten, aber wir müssen uns darauf einstellen. Und der zweite Schritt ist dann Superintelligenz. Das ist eine KI, die sich selbst verbessern kann. Und hier hat Google Deepmind ein Paper veröffentlicht, in dem sie über die Zeit bis 2030 nachdenken. Dort steht drin, dass es keinen Grund gibt, warum es bei menschlichem Niveau aufhören sollte. Der Mensch wird dann zur zweitintelligentesten Spezies auf dem Planeten. In manchen Bereichen zumindest. Das ist eine unfassbare Disruption, und ich glaube, dass wir als Journalisten und Politikerinnen eine Verantwortung haben, das ernst zu nehmen.

Interviewer*in: Was ist das Motiv dieser Firmen und Hersteller, uns und sich selbst zur zweitintelligentesten Spezies zu machen?

Interviewee: Es gibt mehrere Motive. Das eine ist ein wirtschaftliches Interesse. Das Unternehmen, das als Erstes eine allgemeine künstliche Intelligenz entwickelt, hat die größte wirtschaftliche Gewinnchance der Geschichte, weil du mit dieser Technologie dann jeden Markt aufwirbeln kannst und unvorstellbar viel Geld damit verdienen kannst. Wer diese Künstliche Intelligenz kontrolliert, kontrolliert potenziell die gesamte Wirtschaft. Das ist in jedem Fall ein Antrieb, und jeder, der was anderes sagt, lügt. Es gibt dann aber Mitarbeiter und Mitarbeiterinnen in diesen Labs, die das nicht primär aus ökonomischem Interesse machen, sondern, weil sie an diese Technologie glauben. Sie glauben, dass diese Technologie Krebs heilen wird, die Klimakrise lösen wird und andere existenzielle Menschheitsprobleme. Sie sagen, durch KI wird die Welt besser, und deswegen will ich Verantwortung übernehmen. Und das dritte Motiv hat natürlich eine machtpolitische Dimension. Deswegen sind wir jetzt in diesem Rennen zwischen den USA und China um Künstliche Intelligenz. Weil wenn eine Allgemeine Künstliche Intelligenz da ist, wird es natürlich militärische Anwendungen geben. Wird es Möglichkeiten geben, diese Technologie zum Machtgewinn zu nutzen. Und ob die USA zuerst dort sind oder China, wird einen riesigen Unterschied machen. Das ist wie das Rennen zum Mond im Kalten Krieg oder das Manhattan Project im Zweiten Weltkrieg zum Bau der Atombombe. Das ist ein Rennen zwischen Staaten um eine sehr, sehr mächtige Technologie.

Interviewer*in: Bei der Atombombe war es offensichtlich, wieso es wichtig war, Erster zu sein. Warum es so wichtig für Staaten und Unternehmen, im Rennen um Superintelligenz Erster zu sein? Kann es hier nur einen Sieger geben?

Interviewee: Das Unternehmen, das als Erstes die Superintelligenz schafft, wird viele Vorteile haben. Es wird das als Erstes an Kunden verkaufen können, das ist das ökonomische Interesse. Vor allem aber: Es wird einen riesigen Vorsprung beim Sammeln weiterer Daten haben. Weil wenn mein Unternehmen Superintelligenz hat, dann werden viele Menschen

diese Technologie nutzen, und ich kann alle diese Nutzungen verwenden, um das KI-Modell weiterzutrainieren. Und das ist ein sich selbst verstärkender Kreislauf. Ich habe wahrscheinlich so eine Winner-takes-it-all-Dynamik am Markt. Wenn OpenAI das erste Unternehmen ist, das diese Intelligenz hat, dann wird alles hinströmen zu OpenAI. Nutzer werden zu OpenAI kommen. Genauso aber natürlich auch noch mehr Investoren. Und dann gibt es auch einen philosophischen Punkt. Wenn ich das erste Unternehmen bin, das künstliche Superintelligenz entwickelt, habe ich die Möglichkeit, die Werte dieser Technologie zu definieren. Wie schätzt sie Fakten und Unwahrheiten ein? Wofür steht sie? Steht sie für die Menschenrechte? Steht sie für amerikanische Werte, den freien Markt, individuelle Freiheiten? Steht sie für europäische Werte? Steht sie für chinesische Werte? Das alles sind definierende Fragen, weil wenn das eine Technologie ist, die in ganz viele Unternehmen hineingeht, macht es einen riesigen Unterschied, welche Werte sie vertritt. Und insofern ist der, der als Erster dort landet, dann in einer Position, noch schneller weiter vorauslaufen zu können, in ökonomischer Hinsicht, in politischer Hinsicht. Es gibt schon jetzt eine unfassbare Machtkonzentration in der Tech-Branche. Wer sonst kann es sich leisten, diese KI-Modelle zu trainieren? Die hoffnungsvolle Seite davon ist allerdings, dass es momentan nicht so ausschaut. Die letzten Monate zeigen uns eher, dass wir in einem recht pluralistischen Wettlauf sind zwischen Unternehmen, die jede Woche sich um ein paar Prozent weiter übertreffen bei ihren KI-Modellen. Heute ist Google Gemini das stärkste KI-Modell. Das könnte in zwei Wochen schon wieder anders sein. Also momentan schaut es nicht aus, als würde ein Player gewinnen. Aber das könnte sich ändern, wenn ein großer Durchbruch passiert.
Interviewer*in: Spielt Europa eine Rolle in diesem Rennen?
Interviewee: Nein.
Interviewer*in: Kann Europa eine Rolle spielen?
Interviewee: Ich glaube, Europa kann eine Rolle spielen, und das lernen wir beispielsweise aus DeepSeek. Das ist ein KI-Modell, das in China entwickelt wurde, und zwar, wenn man es glaubt, relativ kostengünstig. Der große Durchbruch hier war ein Forschungsdurchbruch, und das könnte in Europa genauso gelingen. Momentan gibt es in Europa eine KI-Firma, die für Aufmerksamkeit sorgt, und das ist Mistral in Frankreich. In diesem Wettrennen zwischen den USA und China nimmt die niemand so ganz ernst. Das könnte aber anders sein, wenn die einen Forschungsdurchbruch haben und wenn sie auf einmal ein sehr starkes Modell hätten, vielleicht das stärkste Modell am Markt. In der Sekunde, wo das passiert, glaube ich, würde sich die Dynamik sehr verändern. Und ich denke, wir sollten stark dafür eintreten, dass das passiert. Gerade wegen der Wertefrage. Wenn wir in eine Zukunft gehen wollen, in der wir pluralistische KI-Systeme haben, in der es nicht die eine große Künstliche Intelligenz gibt, sondern in der Europa möglicherweise auch eine Super-Intelligenz, die europäische Werte vertritt, hat, dann müsste man jetzt alles daransetzen, dass man mehrere kompetitive Start-ups hat, die hier in diesem Rennen mit dabei sind. Denn ansonsten, glaube ich, werden wir ganz was Ähnliches erleben wie bei Social Media, dass es große Firmen gibt, viele hier im Silicon Valley, die den Markt dominieren werden.
Interviewer*in: Angenommen, OpenAI, Mistral oder DeepSeek entwickeln eine KI, die so intelligent ist, dass sie sich selbst immer weiter verbessert und zu einer Superintelligenz wird. Ist so eine Superintelligenz überhaupt noch einholbar, oder ist das Rennen dann gelaufen?
Interviewee: Wir sind hier im sehr spekulativen Bereich. Bei OpenAI gab es ein "Super-Alignment-Team". Das hatte die Aufgabe, Rahmenbedingungen zu entwickeln, wie man Superintelligenz kontrollieren kann, sodass sie zum Wohle der Menschheit agiert. Und das wird nicht einfach. Die großen KI-Labs veröffentlichen immer, wenn sie ein neues Modell machen, einen Testbericht rundherum, wo drinsteht, was sie gefunden haben bei den

Sicherheitstests. Und wir sehen bei den neuen Modellen immer öfter, dass diese KI-Modelle unter gewissen Umständen versuchen, den Nutzer zu täuschen. Unter gewissen Umständen versuchen sie, eine Kopie von sich selbst zu erzeugen und die irgendwo anders hinzulegen, wo die menschlichen Forscher nicht darauf zugreifen können. Das funktioniert noch nicht, und es passiert auch eher selten. Aber dass es passiert, zeigt, dass wir hier eine echte Aufgabe vor uns haben, die, soweit ich das wahrnehme, überhaupt nicht gelöst ist. Und es gibt eine KI-Sicherheits-Community hier im Silicon Valley, die fordert, dass diesem Thema mehr Beachtung geschenkt wird. Ich war vor ein paar Wochen bei einer Demo vor dem OpenAI-Headquarter, wo dazu aufgerufen wurde, die Entwicklung zu stoppen, weil es zu gefährlich sei, das weiterzuentwickeln.

Interviewer*in: Eine Science-Fiction-Vorstellung ist, dass in Zukunft eine gottähnliche KI die ganze Welt steuern und kontrollieren wird. Bei diesem Konzept gibt es zwei Seiten zu bedenken. Einerseits könnte man sagen, eine KI, die es besser weiß als die Menschen, die uns besser regiert als unsere jetzigen Politiker, die fairer ist als unser jetziges Rechtssystem, ist im Sinne der Allgemeinheit. Dann gibt es die andere Seite: Wer garantiert uns denn, dass so eine Super-KI, die uns kontrolliert, tatsächlich unsere Interessen verfolgt? Wie garantieren wir, dass wir keine psychopathische Super-KI entwickeln, die nur ihre eigenen Interessen verfolgt?

Interviewee: Ja, wir wollen keine psychopathische Super-KI. Anthropic hat dazu einen Verfassungskonvent für ihre KI gemacht. Man hat Menschen versammelt und darüber diskutiert, welche Werte die KI vertreten soll. So könnte man diese Werte demokratisch in ein KI-System einbauen. Der Haken ist, dass es immer Leute geben wird, die nicht in der Mehrheit sind, aber wahrscheinlich auch gute Argumente haben.

Interviewer*in: Ich sehe schon die Sollbruchstelle: Es kommt zu einem Machtwechsel in den USA, wo die meisten dieser Firmen sind, und plötzlich heißt es, liberale Werte sind abgeschafft. Es zählt jetzt nur noch radikaler Konservatismus, plötzlich sind die Werte im Sinne der Allgemeinheit dahin.

Interviewee: Wir sehen es jetzt schon bei den Social-Media-Firmen. Facebooks Faktencheck ist plötzlich weg. Mehr maskuline Energie will Mark Zuckerberg bei Meta. Wenn irgendwo Diversität steht, streichen sie es raus aus ihrer Webseite. Weil diese Unternehmen ja einen Anreiz haben, der Politik zu folgen. In einer Art, wie es Journalismus nicht hat, wo es um die Kontrolle von Politik geht. Aber wenn unsere Zukunft die ist, dass jedes Mal, wenn die US-Regierung wechselt, unser gesamtes Wertesystem global durch eine KI gewechselt wird, dann ist das ein riesiges Problem.

(Zsolt Wilhelm, 11.5.2025)

Zur Person

Patrick Swanson (35) ist Unternehmer in San Francisco und hat sich auf den Einsatz von Künstlicher Intelligenz in Medien spezialisiert. Vor seinem Studium in Stanford war er Head of Social Media beim ORF.