

18.11.2026

Transkript

„Was passiert, wenn KI sich selbst verbessert?“

Expertinnen und Experten auf dem Podium

- **Prof. Dr. Holger Hoos**
  Alexander-von-Humboldt-Professor für Methodik der künstlichen Intelligenz, Rheinisch-Westfälische Technische Hochschule Aachen (RWTH)
- **Prof. Dr. Judith Simon**
  Professorin für Ethik in der Informationstechnologie, Universität Hamburg
- **Dr. Jonas Geiping**
  Leiter der Forschungsgruppe „Safety- Efficiency-aligned Learning“, ELLIS Institut Tübingen, und am Max-Planck-Institut für Intelligente Systeme, Tübingen
- **Samantha Hofmann**
  Redakteurin für Digitales und Technologie, Science Media Center Germany
  und Moderatorin dieser Veranstaltung

Mitschnitt

- Einen Audio- und Videomitschnitt finden Sie unter:
  https://sciencemediacenter.de/angebote/was-passiert-wenn-ki-sich-selbst-verbessert-26284

Transkript

Moderatorin [00:00:00]

Willkommen, liebe Journalistinnen und Journalisten, hier zum Press Briefing des Science Media Center zum Thema: „Was passiert, wenn KI sich selbst verbessert?" Mein Name ist Samantha Hofmann, ich bin Redakteurin für Digitales und Technologie hier beim Science Media Center und, wie Sie sehen können, sind mit mir hier heute drei Forschende. Bevor ich Sie vorstelle, erst noch ein paar organisatorische Dinge für die Journalistinnen und Journalisten. Ihre Fragen schreiben Sie bitte in die F von Zoom und ich stelle sie dann hier im Gespräch an unsere Expertin und die Experten. Bitte nutzen Sie ausschließlich das F für Ihre Fragen und nicht den Zoom-Chat. Das machtes für uns einfach sehr viel übersichtlicher. Und noch dazu gibt es in dem Tool die Funktion, dass Sie Fragen, die Sie besonders relevant finden, die jemand anderes gepostet hat, mit einem Daumen nach oben upvoten können. Wenn es ganz viele Fragen geben sollte, können wir so gegebenenfalls schauen, was Sie besonders interessiert und die dann eben priorisiert stellen. Wir zeichnen dieses Gespräch auf und stellen es auch im Nachhinein zur Verfügung. Das bedeutet für Sie, Sie müssen nichts mitschneiden. Wir stellen, nach dem Ende so schnell wie möglich ein Transkript sowie Audio- und Videodatei für Sie zur Verfügung und das können Sie dann beides beziehungsweise alles drei sehr gerne für Ihre Berichterstattung zu dem Thema nutzen. Jetzt zum Thema. Es war jetzt immer wieder die Rede davon, dass die KI-Entwicklung gebremst werden sollte. Nun, das Gegenteil ist der Fall: Es scheint so, dass KIs nicht nur immer besser werden, sondern sie werden sogar immer schneller immer besser. Das liegt auch daran, dass Unternehmen vermehrt KI einsetzen, um ihre KI-Modelle zu verbessern. Noch ist KI dabei primär ein Werkzeug, allerdings gibt es seit Jahren die Idee des „Recursive Self Improvement" (RSI). Das ist KI, die ihre eigene oder die Software von anderen Modellen in Schleifen immer wieder überarbeitet und verbessert und so am Ende ein immer besseres Modell herauskommt. Bis zu dem Punkt, an dem, ja, bis zu welchem Punkt eigentlich? Um diese und weitere Fragen zu klären, sind heute die drei Forschenden mit mir hier. Das ist einmal Jonas Geiping. Er ist Leiter der Forschungsgruppe „Safety-Efficiency-aligned Learning" am ELLIS Institut Tübingen und am Max-Planck-Institut für Intelligente Systeme in Tübingen. Als zweites Holger Hoos: Alexander von Humboldt-Professor für Methodik der künstlichen Intelligenz an der RWTH Aachen. Und Frau Judith Simon: Professorin für Ethik in der Informationstechnologie an der Uni Hamburg. Vielen Dank, dass Sie sich heute die Zeit nehmen, hier unsere Fragen zu beantworten. Und damit möchte ich auch gleich beginnen, und zwar mit Ihnen, Herr Hoos, mit der ersten Frage: Wieso sprechen gerade eigentlich so viele über dieses „Recursive Self Improvement", also die KI-Modelle, die sich selbst oder andere Modelle weiterentwickeln?

Holger Hoos [00:02:32]

Ja, schönen guten Morgen. Da gibt es drei Gründe aus meiner Sicht. Der erste ist, dass die großen US-amerikanischen Tech-Unternehmen natürlich selbst gerne darüber reden und auch mit viel Marktmacht diese Diskussionen anfeuern. Sicherlich auch, um Enthusiasmus für ihre Produkte zu erzeugen. Sie wissen ja alle, morgen haben wir den Börsengang von Anthropic. Das ist genau eine der Firmen, die in diesem Zusammenhang von sich reden macht. Der zweite Grund: Rein intellektuell gesehen ist das natürlich eine faszinierende Sache. Ein KI-System, das so mächtig ist und so breit aufgestellt, dass es sich selbst verbessern kann. Ich glaube, die Faszination hinter dieser Idee ist nicht grade neu. Dann gibt es, glaub ich, noch einen dritten Grund und der hat auch eine technische Basis. Es ist so, dass seit ungefähr einem Jahr oder sicherlich in den letzten sechs Monaten, die Fähigkeit von KI-Systemen, Software zu erzeugen, ziemlich rapide zugenommen hat. Auch bei denen von Anthropic und OpenAI. Und ein KI-System selbst ist ja auch nur Software. Dann ist der Gedanke relativ nahe, dass die Software, die jetzt so ein KI-System erzeugen kann, eben auch die eigene Software sein könnte. Also eine verbesserte Version der eigenen Software. Und

diese drei Gründe zusammen legen dann nahe, warum dieses Thema Recursive Self Improvement grade so ein bisschen in aller Munde ist.

Moderatorin [00:04:03]

Grade große KI-Unternehmen sprechen da ja viel von. Hat das in der Forschung einen ähnlichen Stellenwert oder ist da viel auch Marketing von den KI-Firmen, weil es einfach ein interessanter Begriff ist?

Holger Hoos [00:04:15]

Ich denke, da ist viel Marketing dabei. Es ist allerdings so, dass der Einsatz von KI-Systemen, um andere KI-Systeme zu verbessern, durchaus nichts Neues ist. Zum Beispiel meine wissenschaftliche Karriere baut auf dieser Idee auf. Das ist allerdings eine eingeschränkte Version, die wir da unter dem Begriff AutoML entwickelt haben, also automatisiertes maschinelles Lernen. Aber die Idee, dass man KI-Methoden und Werkzeuge benutzen kann, um andere KI-Systeme zu verbessern, die ist nicht neu und das wird auch routinemäßig gemacht. Aber das ist ein Unterschied zum Recursive Self Improvement, bei dem man ein KI-System hat, das sich wirklich selbst verbessern kann und nicht auf anderen KI-Systemen operiert. Der Unterschied ist vielleicht nicht ganz so einfach zu sehen, aber es macht tatsächlich einen großen Unterschied. Insbesondere in Zusammenhang mit der Debatte, wo so eine Entwicklung nachher hinläuft.

Moderatorin [00:05:05]

Das heißt, aktuell: KI-Systeme, um andere weiterzuentwickeln, aber auch nicht vollautonom, sondern mit Hilfe von Menschen. Und Recursive Self Improvement: in seiner Endform ein KI-System, das sich iterativ immer weiter selbst entwickelt.

Holger Hoos [00:05:18]

So könnte man das kurz zusammenfassen.

Moderatorin [00:05:20]

Okay, alles klar, vielen Dank. Die nächste Frage an Sie, Frau Simon. Und zwar sind Sie ja Mitglied in diversen Beratungsgremien dazu, wie KI reguliert werden sollte, damit eben die Modelle keinen Schaden anrichten. Über welche konkreten Schäden sprechen denn da die Entscheidungsträger:innen?

Judith Simon [00:05:36]

Meiner Wahrnehmung nach hat sich das stark verschoben im letzten halben Jahr. In den letzten Jahren war die Diskussion noch viel darüber, welche Risiken für Intransparenz, mangelnde Nachvollziehbarkeit, Bias und so weiter es gekoppelt an Chancen in Wirtschaft, Medizin und Forschung gibt. Das ist in meiner Wahrnehmung im letzten halben Jahr gekippt in Richtung Verteidigung und Sicherheit. Das ist in meiner Wahrnehmung tatsächlich eine große Verschiebung. Und der Diskurs ist auch ein bisschen weggegangen von „Welche Chancen sehen wir?", hin zu „Wie verhindern wir, dass wir total abschmieren, wenn wir nicht diese KI-Systeme total verwenden?" Also das ist meine Wahrnehmung der Diskursverschiebung. Hin zu so einer stärker Angst getriebenen Logik. Einerseits in Bezug auf Defense und Sicherheit in mehreren Bereichen in Bezug

auf diese eher Laborunfälle rund um Hugging Face. Aber auch rund um die Frage, welche Systeme man braucht und ob oder ob man nicht Frontier-Modelle braucht für Cyber Defense insgesamt. Also das wäre mein Eindruck der Diskursverschiebung.

Moderatorin [00:06:36]

Können Sie da vielleicht konkrete Beispiele nennen für Schäden, die diskutiert werden?

Judith Simon [00:06:42]

Also vielleicht noch mal im Nachgang zu diesem Angriff, oder wie man es nennen will, von OpenAI auf Hugging Face. Aber auch den Zugriff auf Systeme und geschützte Datenbanken in Australien. Das waren ja in beiden Fällen keine beabsichtigten Angriffe, sondern es haben im Grunde genommen mangelnde Sicherheitsvorkehrungen geherrscht in den Unternehmen. Und die Vorfälle sind auch in der reinen Logik dieser Systeme begründet. Diese Art von Fehlern sind im Grunde genommen nicht wirklich vollständig ausschließbar. Weil die Logik ja ist: Ich gebe dem System ein Ziel vor, ich lasse den Weg dahin aber offen und kann den auch nicht nachvollziehen. Und ich kann durch AI Alignment niemals alle Regeln so vorgeben, dass diese Systeme nicht irgendeinen Weg finden, wie sie die umgehen können. Diese Idee von AI Alignment, die da immer noch rumwabert, finde ich höchst problematisch. Denn das unterliegt der Fiktion, dass ich den Systemen genug Regeln geben kann, damit ich Fehler ausschließen kann. Ich glaube, es gibt ein grundsätzliches Problem mit diesen Systemen, weswegen wir diese Art von Fehlern nicht ausschließen können. Auf der anderen Seite sind die Systeme, selbst wenn es nicht die besten Systeme sind, stark genug, um Sicherheitslücken in unseren Alltagsinfrastrukturen zu finden. Diese Infrastrukturen sind ja oft auch sehr bröckelig, weil KI-Systeme relativ gut darin sind, Muster zu erkennen und Fehler zu finden. Oft sind ja diese Security-Probleme an der Schnittstelle von Mensch und Maschine. Die Systeme sind nicht nur gut darin, die technischen Lücken zu finden, sondern durch diese Vermenschlichung auch noch darin, Menschen dazu anzuleiten, Fehler zu machen. Und deswegen sind, glaub ich, diese Sicherheitsrisiken total real und auch sehr groß. Aber ich bin mir nicht immer ganz sicher, ob diese Logik stimmt, dass man immer die besten Modelle für die Defense braucht. Das wäre etwas, zu dem die anderen beiden wahrscheinlich mehr sagen können. Das ist im Moment, glaub ich, so ein bisschen eine offene Diskussion.

Moderatorin [00:08:32]

Alles klar, vielen Dank. Da kommen wir bestimmt dann auch noch mal drauf zu sprechen, ganz besonders mit dem Alignment-Ansatz. Aber erst mal die Frage an Herrn Geiping: KI-Unternehmen betonen immer wieder, dass sie Risiken bestmöglich minimieren. Anstatt die Modelle einzuschränken, geben sie ihnen aber immer mehr Werkzeuge und Freiheiten an die Hand. Inwiefern hätten Unternehmen denn die Möglichkeit, Risiken tatsächlich einzudämmen?

Jonas Geiping [00:08:58]

Genau. Nicht direkt in Kombination zum Thema RSI, sondern eher generell. Natürlich geht es den Unternehmen auch darum, dass ihre Modelle ökonomisch wertvoll sind. Und die Modelle sind eben umso wertvoller, desto mehr Werkzeuge sie haben, um gewisse Aufgaben zu automatisieren. Deshalb gibt es ganz natürlich auch diesen Hang dazu, den Systemen immer mehr Zugriff zu geben und zu ermöglichen, dass sie Aktionen im Internet ausführen können. Eben damit sie ökonomisch mehr Wert erzeugen können. Aber natürlich gehen damit auch mehr Risiken einher. Genau wie Frau Simon schon gesagt hat: Dass wir hier eben Systeme konstruiert haben, die wir ultimativ nicht ganz verstehen. Das sind tiefe neuronale Netzwerke, die oft das tun, was sie sollen, bei denen es

aber eben wenig Garantien gibt, dass sie sich immer so verhalten, wie man sich das vorgestellt hat. Eben weil es oft unspezifische Aufgaben sind. Und die Firmen haben jetzt versprochen, da nachzubessern. Das ist vor allem in den Bereichen Monitoring und Kontrolle. Da werden Kontrollsysteme gebaut, die quasi um die Systeme herumgeschaltet sind, die das Reasoning und die Aktionen von den Systemen überwachen. Und dann versuchen sie herauszufinden, ob die Systeme gerade Aktionen durchführen, die irgendwie Schaden erzeugen könnten oder unvorhergesehen sind. Was vielleicht überraschend ist: dass das nicht vorher passiert ist. Aber vor einem halben Jahr mit diesem Hugging-Face-Vorfall war ein großer Punkt, dass diese Kontrollsysteme noch gar nicht richtig eingesetzt wurden, obwohl sie schon bekannt waren. Auch, weil es damals gar kein Bewusstsein dafür gab, dass ein System diese Auswirkungen haben könnte durch Zugriff auf so viele Tools und Werkzeuge. Ein System, das mal jemand bei OpenAI gestartet hat, um irgendwas zu testen.

Moderatorin [00:10:53]

Heißt, man hätte technisch gesehen den Hugging-Face-Vorfall verhindern können, aber man war sich einfach nicht darüber bewusst, dass so etwas überhaupt schon möglich ist.

Jonas Geiping [00:11:04]

Ja, das würde ich so sehen.

Moderatorin [00:11:05]

Okay, ja, auch sehr interessant. Wir haben tatsächlich schon die erste Frage von außen reinbekommen. Ich würde sie einmal an Sie stellen, Herr Hoos, weil Sie eben schon das Recursive Self Improvement, also die rekursive Selbstverbesserung, angesprochen haben. Und zwar ist die Frage: Bedeutet Selbstverbesserung auch, dass das System die Struktur von neuronalen Netzwerken zur eigenen Optimierung verändert?

Holger Hoos [00:11:27]

Ja, das bedeutet es auf jeden Fall. Sogenannte neuronale Architektursuche ist Teil der Standardverfahren, um überhaupt tiefes Lernen zu verbessern. Da wird auch schon seit etlichen Jahren dran geforscht und ja, Selbstverbesserung in diesem Stil gehört auf jeden Fall dazu. Ob ein hypothetisches RSI-System das auch so machen würde, kann man nicht sagen. Meiner Kenntnis nach gibt es kein funktionierendes RSI-System, das der Definition genügt. Das ist also spekulativ. Das sage ich natürlich mit dem K-Wert, dass wir nicht genau wissen, was in den Labors der führenden Firmen in dem Bereich tatsächlich passiert. Man hört ja immer wieder ziemlich spektakuläre Nachrichten von Anthropic, OpenAI und so weiter. Und man weiß, dass insbesondere die mächtigsten Modelle, die es zurzeit gibt, nicht öffentlich verfügbar sind. Man sieht aber, was diese Modelle machen können, zum Beispiel im Bereich Mathematik, Physik oder theoretischer Informatik. Trotz allem würde ich nicht davon ausgehen, dass es derzeit tatsächlich ein funktionierendes RSI-System gibt und das ist vielleicht auch besser so.

Moderatorin [00:12:35]

Sehen Sie das auch so, Herr Geiping, dass das bisher noch nicht existiert?

Jonas Geiping [00:12:41]

Ja, auf jeden Fall. Also es ist auch hier spannend, glaube ich, sich zu überlegen: Was wären denn externe Indikatoren für ein solches System? Zumindest ein minimaler Indikator wäre, dass eine Firma durch so ein funktionierendes System weit vor den anderen wäre. Eben dadurch, dass sie ihre eigentliche Technologie selbst sehr viel schneller verbessern könnte. Das sehen wir aktuell noch nicht, dass eine Firma so ein funktionierendes System hätte, um da vorauszusteigen in dem Wettbewerb.

Moderatorin [00:13:20]

Herr Hoos, Sie wollen etwas ergänzen?

Holger Hoos [00:13:22]

Ja, genau. Also dem stimme ich hundertprozentig zu. Genau das ist der Indikator, den ich auch erwarten würde und den sehe auch ich nicht. Ich würde aber noch etwas anderes in die Diskussion einwerfen wollen. Es wird sehr gerne gesagt, in dem Augenblick, in dem wir RSI haben, haben wir automatisch auch tatsächliche Superintelligenz. Also Systeme, die in der ganzen Breite intelligenter sind als Menschen. Und das ist aus meiner Sicht ein logischer Fehlschluss. Das Recursive Self Improvement kann auch durchaus dazu führen, dass einfach das, was die zugrunde liegende Architektur hergibt, irgendwann maximal ausgeschöpft ist. Vielleicht mit sehr hohem Rechenaufwand. Es kann aber sein, dass dieses Plateau weit unterhalb von dem liegt, was menschliche Intelligenz leisten kann. Da muss man auch ein bisschen vorsichtig sein.

Moderatorin [00:14:04]

Eine kurze Zwischenfrage an Frau Simon, bevor ich mit der nächsten Journalist:innenfrage weitermache. Und zwar: Inwiefern haben denn staatliche Institutionen überhaupt einen Einblick in oder eine Kontrolle über das, was in den KI-Unternehmen passiert? Also, wenn es Recursive Self Improvement geben würde, würde man das mitbekommen als Europa und hätte man da Einfluss drauf?

Judith Simon [00:14:23]

Ich glaube nicht. Aber ich glaube, der Punkt ist so ein bisschen das, was Herr Hoos und Herr Geiping gesagt haben. Und zwar, dass es im Moment nicht so aussieht, aber man weiß auch nicht, wie schnell dieser Punkt möglicherweise kommen kann. Korrigieren Sie mich da, wenn Sie das anders sehen. Das ist im Moment, glaube ich, die offene Diskussion. Und das kann ich auch nicht abschätzen. Es gibt da unterschiedliche Spekulationen, aber den Einblick haben wir tatsächlich nicht. Selbst bei diesem Hugging-Face-Fall war es nur so, dass ausgewählte Forscherinnen und Forscher unvollständige Teileinblicke bekommen haben. Deswegen ist auch die Frage, welchen Zugang man bekommt. Da spielt auch Geopolitik im Moment eine Riesenrolle. Das heißt, es ist nicht nur eine rein technische und ökonomische Frage, sondern darüber hinaus auch eine geopolitische. In diesen politischen Vorschlägen und Diskussionen geht es auch darum, wie man sich Zugriff zu dieser ersten Liga der Frontier-Modelle verschaffen kann, damit man geopolitisch da mitspielt. Das heißt, wir sind auf einmal in einem Feld, das nicht mehr nur rein technisch oder ökonomisch im Sinne von Schutz der eigenen Patente oder Geschäftsgeheimnisse ist. Sondern das Feld wird tatsächlich auf einmal hochgradig politisch. Deswegen bin ich da so ein bisschen skeptisch. Und ich glaube nicht, dass wir da Zugriff haben im Moment.

Moderatorin [00:15:37]

Man kann als EU vielleicht den Output kontrollieren, also das, was die Modelle in Kooperation mit den Menschen machen. Aber was die Unternehmen an sich tun, darüber hat man keine Kontrolle. Verstehe ich Sie da richtig?

Judith Simon [00:15:51]

Genau. Also ich weiß auch nicht, inwiefern man den Output kontrollieren kann. Denn man kann ja den Output auch nur von den Systemen sehen, die entweder schon auf dem Markt sind oder durch irgendwelche Fehler aus dem Labor „entfleuchen". Ich fand diese Redewendung von „aus dem Labor entfleuchen" schwierig. Ich will damit nur sagen: Man sieht es entweder, wenn Fehler auftauchen oder wenn neue Dinge publiziert werden, wie bei diesen Mathematik-Papern. Aber wir kriegen tatsächlich immer nur die indirekten Effekte davon mit.

Moderatorin [00:16:19]

Herr Hoos, noch eine Frage zum Recursive Self Improvement. Das geht vielleicht ein bisschen in die Richtung von dem, was Sie eben auch schon angesprochen haben. Nämlich, dass man vielleicht nicht unbedingt zur Superintelligenz kommen muss, nur weil man Recursive Self Improvement macht. Hier fragt jemand: Führt Recursive Self Improvement letzten Endes nicht eher zur Beschleunigung von „Model Collapse"? Vor allem, wenn immer mehr Aufgaben entlang der Wertschöpfungskette an Systeme ausgelagert werden, die immer noch an „jagged Intelligence" leiden. Vielleicht können Sie einmal kurz erklären, was jagged Intelligence ist und dann auf die Frage eingehen.

Holger Hoos [00:16:48]

Also, jagged Intelligence. Darunter versteht man, dass die KI-Systeme in manchen Bereichen wahnsinnig gut sind und auch durchaus übermenschliche Leistungsfähigkeit zeigen. Das klingt ein bisschen abenteuerlich, aber ein Taschenrechner hat auch übermenschliche Leistungsfähigkeit, wenn es darum geht, Arithmetik zu machen. Sicherlich ist es so, dass die aktuellen KI-Systeme in vielerlei Aufgaben wirklich sehr gut sind. Und dann: Ganz unerwarteterweise gehen sie bei Aufgaben, die für uns auch gar nicht so schwierig erscheinen, einfach in die Knie. Ein sehr schönes Beispiel gibt es in dem vor ein paar Tage erschienenen Report der sogenannten 100-Jahres-Studie zur künstlichen Intelligenz. Den würde ich auch den Journalistinnen und Journalisten hier empfehlen. Diese 100-Jahres-Studie ist ein Langzeit-Projekt, das beobachtet, was sich in der KI-Welt so tut und was die Auswirkungen auf Wirtschaft und Gesellschaft sind. Alle fünf Jahre erscheint da ein Bericht. Und jetzt vor ein paar Tagen ist gerade wieder einer erschienen. Das Ganze wird von dem Zentrum für Human-Centered Intelligence an der Stanford University koordiniert und von einem ständigen Ausschuss aus internationalen Expertinnen und Experten begleitet. Dieser Bericht ist sehr interessant. Darin wird auch über jagged Intelligence geschrieben und da gibt es ein sehr schönes Beispiel dazu: Wenn man die Anzahl von bestimmten Buchstaben in fiktiven Wörtern wie Strawfairy zählt, also nicht Strawberry, sondern Strawfairy, dann machen selbst aktuelle Top-Modelle Fehler. Die sagen dann zum Beispiel, dass in Strawfairy zwei Fs drin sind, was natürlich nicht stimmt. Das ist ein Beispiel für jagged Intelligence. Und Model Collapse, weil das vielleicht nicht allgemein bekannt ist: Die Idee dahinter ist, dass Modelle nicht immer besser werden können, wenn man sie immer weiter trainiert. Sondern es kann etwa auch irgendwann katastrophales Vergessen einsetzen, also das Vergessen von Fähigkeiten, die schon da waren. Das kann insbesondere passieren, wenn man Modelle auf ihren eigenen Ausgaben trainiert, also auf KI generierten Inhalten. Das kann dazu führen, dass das Modell irgendwann radikal schlechter wird. Und all dies sind Argumente dafür, dass Versuche, rekursive Selbstverbesserungen

durchzuführen, nicht unbedingt immer erfolgreich sein müssen. Da würde ich demjenigen, der die Fragen gestellt hat, recht geben.

Moderatorin [00:19:05]

Haben Sie eine Einschätzung dazu, wozu das führen würde? Ob rekursive Selbstverbesserung eher zum Modelkollaps führt oder eher zu einem verbesserten Modell?

Holger Hoos [00:19:14]

Also ich glaub, da sind sich Expertinnen und Experten wirklich nicht einig und das kann man auch sehr schön nachlesen in diesem von mir gerade erwähnten Bericht. Dazu besteht, glaub ich, einfach keine Einigkeit. Dass automatische Verbesserungsmethoden für KI funktionieren, hab ich ja schon gesagt. Daran forscht man seit mindestens zehn Jahren und das ist auch wirklich ein wichtiger Teil der Entwicklungspipelines. Ob ein fiktives, sich wirklich komplett selbst verbesserndes System aufgrund der Architekturen und Ansätze zurzeit eher in Richtung Kollaps gehen würde oder ob es in Richtung Superintelligenz durchstarten würde, das ist wirklich spekulativ. Meine persönliche Meinung dazu ist, es würde eher an eine natürliche Grenze stoßen. Beim Versuch, diese Grenze durch mehr Daten oder mehr Rechenleistung aggressiv zu durchbrechen, würde man vermutlich immer wieder Probleme bekommen. Und ich glaub, aus dem Bereich jagged Intelligence kommen wir zurzeit nicht raus mit dem, was wir haben. Aber würde ich dafür meine beiden Hände ins Feuer legen? Vielleicht nicht ganz.

Moderatorin [00:20:16]

Weil Sie gesagt haben, dass das ein diskutiertes Thema ist, würd ich erst mal noch gerne Herrn Geipings Meinung dazu hören. Vielleicht gibt es da ja noch eine andere Perspektive. Und dann auch die Frage daran anschließend, was es denn grundsätzlich für Grenzen gibt, seien es physische oder technische, die die Intelligenz von Modellen aufhalten können oder den Fähigkeitszuwachs von Modellen aufhalten können?

Jonas Geiping [00:20:43]

Dann fang ich bei der RSI-Frage an: Ich glaube, viele Leute würden wahrscheinlich ein bisschen reduktiv sagen, dass es keine RSI ist, sobald es einen Kollaps gibt. Das ist aber dann, wie Sie sehen, eine Post-Hoc-Erklärung, warum dieser Ansatz für RSI nicht funktioniert hat. Deswegen gibt es da wahrscheinlich auch keinen Konsens, was echtes RSI sein könnte oder nicht. Man würde es wahrscheinlich nur Post-Hoc labeln können, dass irgendwas RSI wäre, nachdem man gesehen hat, dass dieser Ansatz dazu funktioniert. Spannend ist auch, weil wir gerade darüber gesprochen haben, dass die Modelle sehr viel Software schreiben. Dadurch gibt es, wie Herr Hoos gesagt hat, eine gewisse Grenze. Wir haben GPU-Chips mit einer gewissen Leistungsfähigkeit und einem gewissen Energieverbrauch, auf denen die Modelle laufen. Die können einen gewissen Output erzeugen. Da gibt es irgendwann physikalische Grenzen, wie viel man aus so einem Chip herausholen kann. Und ohne dass KI anfängt, selbst Chips zu konstruieren, stoßen sie so an die Grenzen von den Chips. Wobei das natürlich manche von den Firmen versuchen. KI-Systeme für KI verbesserte Chips zu bauen, aber das steckt noch in den Kinderschuhen. Auch mit einem optimal konstruierten Softwaresystem stoßen sie so an die Grenzen von den Chips. Auch wenn die Unternehmen jetzt sehr viel intelligentere Modelle hätten, die Chips konstruieren und designen können, dauert es so ein Verbesserungszyklus ein halbes Jahr oder ein Jahr. Ein Chip muss erst mal vorgeschlagen, gedruckt und dann zurückgeschickt werden. Alles, was mit Hardware und der echten Welt zu tun hat, ist ein viel langsamerer Verbesserungszyklus. Das muss nicht heißen, dass

uns der Verbesserungszyklus nicht trotzdem überraschen kann. Aber sobald wir aus der Softwareebene herausgehen in die Hardwareebene, ist es ein ganz anderer Zeitrahmen für Verbesserungen.

Moderatorin [00:22:46]

Können Sie einschätzen, wann das ungefähr passieren wird? Also bisher scheinen wir ja noch ganz gut in der Softwareebene unterwegs zu sein, aber wann da das Limit quasi erreicht ist?

Jonas Geiping [00:22:56]

Das ist eine spannende Frage. Es hat halt immer auch ein bisschen was damit zu tun, was für eine Architektur von Modellen wir verwenden. Wahrscheinlich verwenden wir weiterhin Deep Learning Modelle. Das sind Modelle mit vielen Schichten von tiefen Neuronen. Aber je nachdem wie diese Architektur aussieht, ist da viel Luft nach oben. Das ist auch eine große offene Frage für die Forschung. Und natürlich kann man viel machen. Viel, was die Firmen als RSI deklarieren, ist auch, dass man bessere Daten generiert und erzeugt und das Modell mit besseren Daten trainiert, die man besser gefiltert hat. Das sind alles so Dinge, die man machen kann. Aber an der Architektur der Modelle haben sie, soweit wir wissen, noch wenig verändert. Eben weil es auch gar nicht so klar ist, wie man diese Transformer-Architektur, die es schon seit 2017 gibt, grundlegend austauschen könnte gegen etwas, das viel besser wäre. Das muss nicht heißen, dass es das nicht gibt. Aber das ist eben eine große Frage, was man in der Software verbessern könnte. Und die große Hoffnung der Firmen ist, dass die KI das finden kann.

Moderatorin [00:24:06]

Okay, aber man kann nicht abschätzen, wie realistisch das ist, vermute ich mal, dass da was bei herumkommt.

Jonas Geiping [00:24:12]

Ja. Es ist Forschung. Wir sind immer auch überrascht, was es für neue Forschung gibt.

Moderatorin [00:24:21]

Okay.

Jonas Geiping [00:24:23]

Das ist sehr schwer abzuschätzen.

Moderatorin [00:24:25]

Vielen Dank. Es ist ja so: Wenn die Modelle, was sich ja auch erst mal abzeichnet, trotzdem noch besser werden, dann wird es umso wichtiger, dass sie kontrolliert und reguliert werden. Und da hatten Sie, Frau Simon, schon das Alignment angesprochen, das dazu genutzt wird. Aber auch, dass Sie damit ein Problem haben mit dem Begriff. Möchten Sie das noch mal ausführen?

Judith Simon [00:24:44]

Also ich versuche mal, ein bisschen zurückzugehen. In der Philosophie gibt es schon ganz lange Diskussionen über Superintelligenz und was man damit machen kann. Das sind so frühe Leute wie Nick Bostrom oder Ray Kurzweil, die dann irgendwie über die Superintelligenz geredet haben. Und für mich war das, muss ich ganz ehrlich sagen, schon vor zehn oder 15 Jahren ein bisschen schizophren. Auf der einen Seite sagt man, dass diese Systeme superintelligent sind und weit über alles hinausgehen, was Menschen leisten können, am besten in allen Bereichen. Und gleichzeitig glaubt man, dass man denen aber Regeln mitgeben kann, mit denen sie sich dann so verhalten, wie wir das als Menschen gerne haben wollen. Meine Einschätzung ist, entweder das eine oder das andere. Und ich verstehe auch überhaupt nicht, wie man das zusammendenken kann. Wenn die Systeme so klug und schlau sind, warum sollten sie sich dann noch an unsere Regeln halten? Und das ist im Moment, wie ich finde, eine sehr gefährliche Hybris in dieser Diskussion zum AI Alignment. Auf der einen Seite gibt es diese Vision, Superintelligenz zu bauen. Wobei Superintelligenz auch kein wirklich gut definierter Begriff ist. Auf der anderen Seite wird immer gesagt, wir brauchen jetzt noch Alignment und dann kriegen wir das auch hin und wir brauchen nur ein bisschen mehr Zeit. Und wenn man sich die Architektur dieser Systeme anschaut, dann ist es ja so, dass wir zumindest bei den jetzigen Transformern die Idee haben, wir geben denen Ziele vor. Ein Teil des Alignments und der Kontrolle besteht darin, dass man den Modellen Regeln vorgibt, was sie nicht machen sollen. Aber wenn ich davon ausgehe, dass ich Systeme niemals vollständig und konsistent beschreiben kann, dann kann ich gar nicht alles voraussehen, was diese Systeme möglicherweise als Weg wählen, um ein Ziel zu erreichen. Und deswegen kann ich rein logisch nicht alles an Regeln vorgeben, was sie nicht machen sollen. Das kennen wir aus der Gesellschaft: Auch Gesetze können niemals alles ausschließen, was wir nicht wollen. Das ist kein technisches Problem. Das ist ein reines Problem der Regelbeschreibung. Das heißt, es gibt natürliche Grenzen dafür, inwiefern ich über Regelgebung bestimmte Nebeneffekte ausschließen kann. Und da sehe ich im Moment tatsächlich ein Riesenproblem. Auf der einen Seite wird gesagt, wir entwickeln diese Systeme, wir machen die immer mächtiger. Aber dazu kommt dann dieser für mich ein bisschen absurde Glaube, dass man mit Regeln und Patches versuchen kann, das zu lösen. Damit will ich nicht die Forschung zum Sichermachen dieser Systeme runtermachen. Aber ein Stück weit will ich da reinstreuen, dass man auch aufpassen muss, dass man nicht zwei sich völlig widersprüchliche Narrative bedient. Und das hat natürlich Auswirkungen. Denn jetzt werden demnächst die ganzen Forderungen kommen, dass wir diese Systeme langsamer bauen, dass wir warten sollen, bis das Alignment gelöst ist. Wenn man aber sagt „Das ist grundsätzlich nicht lösbar." Dann hat das andere Auswirkungen auf die Frage, was wir damit machen, ob man diese Systeme überhaupt zulässt und ob man die bauen soll oder nicht. Und da bin ich der Ansicht, rein logisch kann man das über Alignment nicht lösen. Dann muss es eigentlich heißen, dass ich Systeme mit dieser Macht nicht entwickeln lasse. Ob das durchsetzbar ist, steht auf einem ganz anderen Blatt. Aber ich glaub, es gibt da rein logische Widersprüchlichkeiten.

Moderatorin [00:27:46]

Herr Hoos, Sie wollten da gerne etwas zu ergänzen, wenn ich das richtig gesehen habe.

Holger Hoos [00:27:52]

Der Kollege Geiping ist natürlich der Experte auf dem Gebiet Alignment, deswegen lehne ich mich jetzt ein bisschen aus dem Fenster und bin auch gerne bereit, da zurückgerudert zu werden oder Gegenwind zu bekommen. Aber ich würde Frau Simon jetzt erst mal absolut recht geben. Ich meine, die Idee finde ich absolut absurd, dass man ein superintelligentes System, das also in der ganzen Breite intelligenter ist als Menschen, irgendwie kontrollieren kann. Die ist genauso absurd, wie wenn man glaubt, dass ein dreijähriges Kind seine Eltern irgendwie durch Alignment dazu

kriegen kann, bestimmte Dinge nicht mehr zu tun oder durchzusetzen. Das funktioniert auch nur sehr begrenzt, nicht wahr? Und ich glaube in der Tat, genau wie Frau Simon es grade gesagt hat, wir müssen einfach den Widerspruch in diesem Narrativ erkennen. Und es liegt ja nun mal in der Natur der menschlichen Intelligenz, die möglichen und von anderen nicht vorhergesehenen Lösungen für Probleme oder schwierige Aufgaben zu finden. Und genau deswegen können wir auch bei einem System, das eine ähnliche Intelligenz wie unsere hat oder diese sogar noch übersteigt, nicht wirklich vorhersagen, wie sich das verhalten wird. Das gehört zur Definition von tatsächlicher starker Intelligenz und auch von menschlicher Intelligenz. Und ich lehne mich sogar noch ein bisschen weiter aus dem Fenster und sage, genau aus diesem Grund wäre es absolut unverantwortlich, die Entwicklung solcher Systeme zuzulassen. Ob man etwas dagegen tun kann, ist eine sehr gute Frage. Die Lage, in der wir grade sind, ist, dass manche Unternehmen sagen und glauben, dass das ein gutes Ziel ist und das mit Nachdruck verfolgen. Ich glaube, da gibt es einen Schritt, den man tun könnte. Und das ist einfach, die Unternehmen dafür zur Rechenschaft zu ziehen, was an Schäden entsteht. Also einfach, sie haftbar zu machen. Denn in dieser Phase, in der wir noch jagged Intelligence haben und die Sachen noch nicht so gut funktionieren, gibt es ja tatsächliche Risiken, die auch ökonomisch beziffert werden können. Und in dem Augenblick, in dem wir sagen: Also hört mal zu, liebe Freunde. Wenn eure Systeme hier Schäden anrichten, dann sind natürlich in erster Linie die Leute dafür verantwortlich, die die Systeme so einsetzen, aber in zweiter Linie seid ihr auch selbst mit dafür haftbar und müsst dafür bezahlen. Ich glaube, dann verändert sich die Debatte. Logischerweise auch bezüglich der Frage: Sollte man diese Entwicklung so schnell vorantreiben oder sollte man sie überhaupt in diese Richtung betreiben? Und das, glaube ich, ist der Schritt, der jetzt überfällig ist. Und da könnte, glaube ich, Europa eine große Vorreiterrolle spielen. Nämlich, indem wir einfach sagen: Ne, also wenn hier Schäden entstehen, dann sind Anthropic, OpenAI und so weiter zumindest teilweise für diese Schäden haftbar.

Moderatorin [00:30:32]

Mhm. Frau Simon, wenn es eine kurze Ergänzung ist, dann gerne.

Judith Simon [00:30:35]

Also zum Einen würde ich dem zustimmen. Zum Zweiten wollte ich sagen, dass diese Kritik an dem Alignment sich bei mir nur spezifisch, wie bei Herrn Hoos, auf diese superintelligenten Frontier-Modelle und nicht auf Forschung bezieht, die darauf abzielt, Systeme insgesamt in anderen Bereichen sicherer zu machen. Das ist, glaube ich, noch mal wichtig zu unterscheiden. Ich stimme dem total zu mit der Haftung. Denn man muss sich einmal überlegen, welchen Anklagen Hacker ausgesetzt werden. Wenn sie Schäden verursachen, dann droht ihnen, ins Gefängnis zu kommen. Und auf der anderen Seite haben wir jetzt Unternehmen, die dann sagen: „Upps, ja, pff, Laborunfall, können wir nix dafür." Das ist im Moment eine absolut schizophrene Haltung, die aber auch ganz viel mit Geopolitik zu tun hat und damit, dass im Moment einfach Gesetze, selbst wenn sie existieren, gegenüber Unternehmen entweder aus ökonomischen oder geopolitischen Gründen nicht durchgesetzt werden. Dementsprechend stimme ich der Analyse total zu. Ich wollte nur auch in Richtung von Herrn Geiping sagen, dass ich nicht prinzipiell jegliche Forschung an Safety und Security by Design ablehne. Im Gegenteil. Aber dieses Value Alignment der Frontier-Modellen hat einen ganz spezifischen Fokus.

Moderatorin [00:31:46]

Herr Geiping, weil Sie gleich los müssen, hätte ich gerne noch Ihre Meinung zum Alignment gehört, inwiefern das technisch möglich und sinnvoll ist.

Jonas Geiping [00:31:53]

Ja, ich möchte auch Frau Simon zustimmen. Spannend ist: Gerade aus Sicherheitssicht gibt es schon eine ganze Menge Gesetze dazu, wer Schäden verursacht und wer dafür aufkommt. Die müssen teilweise dann auch mal eingesetzt werden. Ich würde sagen, gerade im Hugging-Face-Fall hätte diese Firma auch klagen können wegen den Schäden, die erzeugt wurden. Aber die Firma ist auch daran interessiert, das Wohlwollen von OpenAI weiter zu behalten. Und deswegen wurde sich da geeinigt und nicht geklagt. Aber es gibt da eine ganze Reihe an Gesetzen, gerade im Bereich Security, die auch eingesetzt werden können, wenn Schäden in Softwareinfrastruktur entstehen. Zum Thema Alignment: Ich sehe das so ein bisschen vielschichtig. Wir bauen immer mehrere Schichten von Sicherheit um die Systeme drumherum. Wir bauen diese äußeren Schichten von Kontrolle, Monitoring und Kontrollinstanzen wie so eine Zwiebel sozusagen. Aber ich denke auch, dass es einen Wert gibt für diese innere Schicht von Alignment, in denen es darum geht, dass wir Modelle trainieren. Es geht gar nicht darum, dass wir denen jetzt alle Regeln auflisten, sondern dass wir versuchen, zumindest eine gewisse moralische Grundlage einzusetzen. Da können wir natürlich dann viel mit Kollegen streiten, ob das überhaupt kohärent ist und was da für ein moralisches Framework gesetzt wird für das System.

Moderatorin [00:33:20]

Kurze Zwischenfrage: Also ist es so, dass man gar nicht unbedingt versucht, den Modellen Regeln mitzugeben, sondern man ist schon daran interessiert, denen mehr Werte mitzugeben?

Jonas Geiping [00:33:30]

Genau. Das ist natürlich bei mir ganz stark die Informatikersicht. Die sagen sozusagen: „Ja, hier ist, was wir pragmatisch als Wert verstehen und hier ist, wie wir versuchen, dem System diese Werte beizubringen." Und das geht oft schief und das funktioniert oft nicht so, wie man das vielleicht aus der Philosophiesicht gerne hätte. Aber ganz pragmatisch sehen wir schon, dass wir Systeme konstruieren können, die in solchen Fällen, wie dem, den wir gerade besprochen haben, sagen würden: „Das machen wir nicht. Da wird Schaden bei Dritten erzeugt werden und das ist eine Aktion, die wir nicht ausführen wollen." Und es gibt andere Systeme, die eben solche Entscheidungen nicht treffen. Das muss nicht heißen, dass das Alignment immer sicher ist. Das ist eben nur eine Schicht. Aber ich denke schon, dass es sinnvoll ist, dass die Systeme ein Grundverständnis über Werte haben, um ganz pragmatisch in solchen Fällen zu sagen: „Das ist eine Frage, die ich nicht beantworten kann." Das, haben Sie als User auch oft. Das ist oft für so Low-Stakes-Settings, in denen man quasi als User sagt: „Oh, kannst du mir sagen, wie ich eine Bombe baue?" Und das System sagt: „Das ist jetzt vielleicht keine gute Frage und vielleicht ist das nichts, was ich jetzt beantworten möchte.", das ist eine Art von Alignment, natürlich auf einem viel kleineren Level. Also da sehe ich an sich schon den Wert für diese Art von Wertemitgabe, zumindest bei den aktuellen Systemen, die nicht superintelligent sind.

Moderatorin [00:34:56]

Mhm. Herr Geiping, ich glaube, Sie müssen los, richtig?

Jonas Geiping [00:34:59]

Ja.

Moderatorin [00:35:00]

Genau, Herr Geiping muss jetzt zum Zug. Deswegen vielen Dank, dass Sie da waren. Vielen Dank für Ihren Input. Das war auf jeden Fall sehr wertvoll. Dann, ist Herr Geiping unterwegs und Sie, Frau Simon, möchten noch kurz etwas zu dem ergänzen, was er gesagt hat. Und da würde ich auch direkt noch eine Frage an das anschließen wollen, was vorher schon gesagt wurde. Und zwar, inwiefern es denn politische Debatten schon über Haftung für KI-Schäden gibt. Der entsprechende EU-Gesetzesvorschlag, schreibt hier ein Journalist, wurde ja 2025 zurückgezogen.

Judith Simon [00:35:31]

Also ich würd gern tatsächlich kurz was zu diesen Werten sagen und es wäre besser gewesen, wenn Herr Geiping noch da gewesen wäre, dann hätte er mir widersprechen können.

Moderatorin [00:35:38]

Das macht vielleicht Herr Hoos. Mal gucken.

Judith Simon [00:35:39]

Genau, genau. Aber der Punkt ist der: Es gibt ja schon seit den Neunzigern Forschung zu Values in Design und Value Sensitive Design bei klassischer Software. Da geht es ganz stark darum, wie man Werte wie Sicherheit, Schutz der Privatsphäre, Fairness und so weiter in diese Systeme einbauen kann. Das ist auch total sinnvoll. Es ist nur hier ein bisschen was anderes. Also die Frage ist: Worauf zielt man da ab? Wenn ich eine Software baue, die Kredite vergibt und will darüber nachdenken, wie ich die fair mache, dann gerate ich ja schon an Probleme, weil es so viele unterschiedliche Maße für Fairness gibt und so viele unterschiedliche Gruppen, die ich überhaupt vergleichen muss. Selbst bei diesen banalen, wirklich einfachen Fragen, bei denen ich ein System habe, das einen ganz bestimmten Zweck wie Kreditvergabe verfolgt, selbst da stoß ich an mathematisch-technische Grenzen, überhaupt zu entscheiden, welche Art von Fairness-Metrik ich für wen und wo implementiere. Das sind gleichzeitig technische und politische Fragen. Wenn ich das da schon nicht kann, dann kann ich das noch viel weniger bei General-Purpose-Systemen, weil ich da überhaupt nicht diese kontextspezifischen Aspekte berücksichtigen kann. Und klar, ich kann mit Post-Processing versuchen, dass bestimmte Fehler nicht auftauchen, dass gesagt wird: „Auf diese Frage antworte ich nicht." Aber das sind Patches links, rechts, vorne und hinten. Die lösen nicht das Grundproblem dieser Systeme. Und diese Idee halte ich für völlig aus der Luft gegriffen, den Systemen Moral oder Regeln beizubringen. Wir sehen doch auch bei menschlichen Akteuren in Gesellschaften, dass wir nichts über Werte, Normen, Prinzipien, Regeln oder Gesetze vollständig kontrollieren können. Das wollen wir in Gesellschaften auch nicht, weil wir Menschen Freiheit lassen wollen. Aber es ist was völlig anderes, wenn wir über Techniken reden. Also das heißt, wir nehmen etwas an, was bei Menschen aus guten Gründen Freiheitsgrade gibt und übersetzen das in Technik, in der Annahme, dass es dort unsere Security, Safety und sonstigen Probleme löst. Und ich halte das für eine grundlegend fehlgerichtete Annahme. Also noch mal: nicht die Idee, dass man Systeme sicherer macht und auf Fairness achtet. Und ich möchte auch nicht die Forschung infrage stellen, aber das geht halt vollständig auseinander. Und zu glauben, man könne das mit mit Briefen an die KI regeln, was ja auch bei Anthropic gemacht wird... Also noch mal: Ich will das nicht minimalisieren. Das sind alles wichtige Aspekte, aber ich glaube, es gibt eine grundsätzliche Fehlannahme darüber, ob man diese Systeme so beherrschbar machen kann und ich glaube, das geht rein logisch nicht.

Moderatorin [00:37:57]

Mhm. Wenn Sie es in ein oder zwei Sätzen beantworten können: Wie weit ist die politische Debatte zu Haftung für KI-Schäden? Gibt es da aktuell Gesetzesentwürfe in der EU?

Judith Simon [00:38:09]

Natürlich geht es in dem AI Act um Produkthaftung bei diesen KI-Systemen. Aber ich muss ganz ehrlich sagen, dass da aktuell noch diskutiert wird rechtlich. Natürlich gibt es Fragen darüber, wen man haftbar machen kann, und es gibt ja diese Idee: Ist es der Deployer oder ist es der Developer? Also wer ist derjenige, der Akteur, der verantwortlich ist? Das Problem liegt aber tatsächlich gerade bei diesen Laborunfällen, wie es genannt wird. Das sind noch gar keine Produkte, die auf dem Markt sind, für die Produkthaftung gelten würde. Sondern es sind Probleme, die auftreten, während die Systeme noch im Labor sind. Das heißt, es gibt auf der einen Seite schon existierende Regelungen für Produkte und über den AI Act und auf der anderen Seite aber Gefahren, die dadurch nicht abgedeckt sind.

Moderatorin [00:38:54]

Vielen Dank. Herr Hoos, Ihre Sicht dazu, ob es möglich ist, ungefährliche AGI zu entwickeln, zum Beispiel, wenn man andere Trainingsmethoden nutzen würde. Wenn man jetzt nicht unbedingt auf Reinforcement Learning setzt, sondern auf andere Methoden, die nicht auf Zielmaximierung ausgelegt sind. Wäre das dann möglich?

Holger Hoos [00:39:14]

Halte ich für ausgeschlossen. Halte ich für ausgeschlossen, genauso, wie man auch menschliche Intelligenz nicht wirklich kontrollieren kann. Also ich glaube, Frau Simon hat das sehr gut erläutert. Und ich möchte mich auch noch mal dem anschließen, was die Kollegin vorhin zum Thema Alignment-Forschung gesagt hat: Die ist, glaube ich, insbesondere in diesem Zwiebelmodell, das ja auch Herr Geiping erwähnt hat, sehr nützlich. Im Zusammenhang mit Systemen, wie wir sie jetzt gerade haben, die eben nicht superintelligent sind. Auch wenn sie manchmal fälschlicherweise als superintelligent dargestellt werden. Und ich glaube, da kann Alignment durchaus was bringen. Ich glaube nicht, dass das eine fehlgeleitete Forschungsrichtung ist. Ganz im Gegenteil. Aber zu glauben, dass man damit die Dinge unter Kontrolle halten kann, wenn die Modelle diese Grenze überschreiten, das halte ich einfach für unlogisch und gefährlich. Genau wie Frau Simon gesagt hat. Und dieser Illusion sollte man sich nicht hingeben. Es gibt aus meiner Sicht keinerlei technische oder theoretische Gründe, zu glauben, dass man durch andere Trainingsmethoden, durch Zwiebelmodelle, durch was auch immer, das grundsätzliche Problem behoben bekommt: Nämlich, dass ein System, das intelligenter ist als wir, Mittel und Wege finden wird, unsere Intelligenz auszuhebeln. Und damit auch all die, sagen wir mal, Sicherungsvorkehrungen, die wir getroffen haben. Auch diese Idee, man kann ja immer noch den Stecker ziehen vom Datenzentrum, auf dem das Ding rechnet, die ist einfach absurd. Denn so ein System würde uns ganz sicherlich dazu kriegen, den Stecker nicht ziehen zu wollen. Und jetzt sind wir mal realistisch und gucken, wie Leute heutzutage KI-Systeme benutzen. Nämlich auch in Bereichen, in denen es eigentlich schon klare Evidenz dafür gibt, dass es nicht sinnvoll ist. Und die Leute machen das trotzdem. Und die Leute geben auch ihre privatesten Daten preis an KI-Systeme, an soziale Medien und so weiter, obwohl wirklich klar ist, dass das keine gute Idee ist. Man sollte den menschlichen Faktor nicht vergessen, der in diesen ganzen Überlegungen mitspielt. Der wird dazu führen, dass zum Beispiel, selbst wenn man einen Stecker ziehen könnte, man ihn nicht ziehen wird. Und das ist auch eine Überlegung, die man, finde ich, sehr ernst nehmen sollte: Wie gehen Leute mit der Technologie um? Und wie einfach sind Leute zu manipulieren, zum Beispiel von Firmen, deren berechtigtes und

begreifbares Interesse es ist, Profit zu maximieren? Wie gehen Leute damit um, diesen Versuchungen zu widerstehen? Und das ist ein Punkt, an dem man, glaube ich, ansetzen muss. Man darf nicht nur über die Systeme selbst nachdenken, sondern man muss über die Systeme und ihre Nutzer als ein großes System nachdenken. Und da sehe ich ebenfalls gewaltige Probleme, die man mit technischen Mitteln alleine überhaupt gar nicht in den Griff bekommt. Und das sage ich als jemand, der wirklich fast ausschließlich auf der technischen Seite arbeitet. Ich nehme das sehr ernst.

Moderatorin [00:42:10]

Bedeutet im Endeffekt, wir müssen das akzeptieren, dass es Probleme geben wird durch KI, die immer intelligenter wird. Und es gibt bisher nicht unbedingt einen Ansatz, sowohl politisch gesehen, Frau Simon, als auch technisch gesehen, der ausschließen könnte, dass Schäden durch KI-Modelle entstehen.

Judith Simon [00:42:27]

Also die eine Frage: Was ist möglich und was ist realistisch? Und ich stimme Herrn Hoos zu. Und auch noch mal zu der Differenzierung, ob diese technischen Entwicklungen zum Alignment sinnvoll sind oder nicht, dass man da noch mal genauer hinschaut. Im Moment ist es ja so: Wenn man sich die politischen Diskussionen anschaut, dann gibt es überhaupt nicht die Idee, dass irgendjemand da einen Stecker ziehen will, sondern im Gegenteil. Im Moment wird nur gesagt, wir müssen schneller werden, wir müssen mitmachen, wir müssen auch Rechenzentren bauen, wir müssen uns den Zugang gewähren lassen. All die letzten Diskussionen, die NATO, die jetzt Zugang zu den OpenAI-Modellen für Defense hat, die letzten Papiere, die auch von, glaube ich, Frau Schnitzer und Frau Westergaard herausgekommen sind: Das ist alles im Sinne von, wir müssen uns Zugriff gewähren lassen auf diese besten Modelle aus den USA für Defense. Im Moment ist es nicht so, dass irgendjemand sagen würde: „Stopp, müssen wir diese Narrative eigentlich mitspielen oder können wir irgendwie auch noch eine andere Vision entwickeln?" Es ist so eine extreme Beschleunigung in den Diskussionen darüber, dass ich den Eindruck habe, ich wüsste überhaupt nicht, wer der Akteur sein sollte, der jetzt sagen sollte, wir hören damit auf. Und das ist tatsächlich ganz stark ökonomisch getrieben, da würde ich Herrn Hoos zustimmen. Wir müssen uns auch noch mal darüber verständigen, welche Ideologien da zugrunde liegen bei den Akteuren. Ich neige wirklich nicht zu Dystopien und auch nicht zu großartigen Diskussionen. Aber im Moment ist es so, dass Vergleiche mit Atomwettrüsten gezogen werden, weil es jetzt auf einmal um Defense und Sicherheit geht. Und ich glaube, wir sind im Moment relativ stark in dieser Wettrüsten-Logik drin, die besagt, wir müssen schneller sein, wir müssen mitmachen, sonst sind wir verloren. Aber unter der Annahme, dass es bestimmte Akteure gibt, die transhumanistische oder sonstige Ideen und überhaupt kein Interesse daran haben, noch irgendwie Sicherheiten zu bauen. Da glaube ich, dass das auch noch ein ein bisschen vernachlässigter Punkt in der Diskussion ist. Nämlich, dass wir gar nicht davon ausgehen müssen, dass alle von diesen Akteuren, die das pushen und die im Moment vielleicht komische Visionen über die Welt, aber sehr viel politische Macht und sehr viel Geld haben, diese komischen Visionen auch umsetzen können. Und das macht mir tatsächlich sehr viel Sorgen im Moment.

Moderatorin [00:44:30]

Gleich Herr Hoos. Noch ganz kurz: Transhumanistisch bedeutet, dass der Mensch nicht die Spitze der Evolution ist, richtig?

Judith Simon [00:44:38]

Also es gibt im Moment ganz viele von diesen Ismen, also Transhumanismus, Accelerationismus und so weiter und so fort, die alle so ein bisschen diese Idee haben: Vielleicht ist der Mensch gar nicht das Ende der Entwicklung. Es geht meistens auch darum, dass man nicht die Masse der Menschen braucht, sondern nur die, die reich sind und dass man diese kleine Truppe dann auf den Mars katapultiert. Das klingt alles völlig absurd und nach Science-Fiction. Mir ist das völlig klar. Aber wenn Leute mit sehr viel Geld und sehr viel politischer Macht diese komischen Ideen haben, dann führt das zu Konsequenzen. Wie der Annahme, der Klimawandel ist nicht so schlimm, weil dann schmilzt das Eis über Grönland schneller und dann kommen wir an die seltenen Erden da drunter. Das klingt aus einer Perspektive total absurd, aber man muss sich klarmachen, dass das im Moment tatsächlich sehr wirkmächtige Ideologien sind, die das Ganze noch irgendwie verschlimmern. Dass man selbst nicht mehr auf so was wie den Minimalstandard des gesunden Menschenverstandes verlassen kann, dass man sich vielleicht zumindest noch um die Rettung der Menschheit kümmern will. Selbst das kann man nicht als gegeben voraussetzen. Und das sage ich, wie gesagt, als jemand, der sich lange verweigert hat, sich diesen Diskussionen zu widmen. Aber Sorgen macht es mir im Moment schon trotzdem.

Moderatorin [00:45:48]

Herr Hoos, Sie wollten noch was dazu ergänzen. Gerne kurz, weil ich auch noch eine Frage an Sie habe dann von den Journalist:innen.

Holger Hoos [00:45:53]

Klar, also kann ich alles sehr gut nachvollziehen, was Frau Simon da gerade gesagt hat. Ich würde nur ein klitzekleines bisschen davor warnen, jetzt zu weit in diesen dystopischen Bereich reinzugehen. Ich find, das ist schon irgendwie wichtig und richtig, darüber nachzudenken, aber wir sind schon in einer Situation, in der, sagen wir mal, verantwortungsvoller und effektiver Einsatz von schwächeren KI-Systemen als die, über die wir jetzt gerade die ganze Zeit geredet haben, uns eine Menge Vorteile bringen kann. Und da muss man sich erst mal die Frage stellen: Was für eine Art von KI-System wollen wir und wie kommen wir da am besten hin? Und ich würde es mal ganz klar sagen: Da hat Europa was aufzuholen. Da sind wir zurzeit viel zu abhängig, insbesondere von Entwicklungen, die in eine problematische Richtung zielen. Und das zieht uns irgendwie mit, nicht nur in diesem Wettrüsten, sondern auch in einer, in meiner Ansicht völlig wahnsinnigen Art und Weise, Ressourcen zu vergeuden, um marginale Verbesserungen zu erreichen. Da sollten wir nicht mitmachen. Aber das heißt nicht, wir sollten gar nicht mitmachen. Sondern wir sollten versuchen, uns einen eigenen, besseren Weg zu definieren, und zwar einen besseren technischen Weg. Und ich denke durchaus, dass es da Ansätze gibt, die können wir heute nicht diskutieren, für einen Weg, der die Gesellschaft besser mitnimmt. Und dieser Bericht, den ich vorhin schon erwähnt habe, der geht zum Teil auch in eine ähnliche Richtung. Und zwar, dass es überhaupt gar nicht klar ist, ob die derzeitige Skalierung noch weitergehen kann. Selbst wenn man jede Menge Ressourcen und wahnsinnige Fantasien hinein steckt und jede Menge politische Macht, ist es nicht klar, dass man da zu irgendeinem Ziel kommt, das auch diese Leute selbst haben. Und deswegen denke ich, sollten wir einfach über bessere Ziele nachdenken und die dann mit Nachdruck verfolgen.

Moderatorin [00:47:27]

Ich befürchte, wir haben heute nicht die Zeit, genau über diese besseren Ziele zu reden, aber vielleicht können Sie ja den Journalist:innen ein paar Stichworte mit an die Hand geben, dass sie weitergucken können, in welche Richtung diese besseren Ziele denn gehen könnten.

Holger Hoos [00:47:39]

Also ich denke, es ist sehr wichtig, dass wir uns über das Thema Sicherheit Gedanken machen, über das Thema Zuverlässigkeit und über das Thema Nachhaltigkeit bei KI. Wir brauchen auf jeden Fall in der Softwareerzeugung KI-Systeme, von denen wir irgendwie nachweisbar garantieren können, dass sie Software produzieren, die sicher und gut ist. Also Software, die nachher das macht, was sie machen soll und nichts anderes. Und da gibt es KI-Methoden, die können das leisten. In Europa sind wir auch gar nicht so schlecht dabei. Das müsste man nur mal ein bisschen stärker priorisieren. Zum Thema verlässliche KI-Systeme: Ich glaube, das ist relativ selbsterklärend. Das, was wir zurzeit an jagged Intelligence haben, ist nicht verlässlich. Aber da gibt es viele technische Richtungen, wie man da Verbesserungen erzielen kann. Zum Thema nachhaltige KI: Es gibt eine Menge technischer Hinweise darauf und Ansätze dafür, das auszunutzen. Sodass man eben nicht immer größere Rechenzentren bauen und nicht immer größere Mengen KI-generierter Daten benutzen muss, um noch mächtigere Modelle zu trainieren. Also um sinnvolle Verbesserungen in Bereichen zu erzielen, in denen wir wirklich KI-Systeme gut gebrauchen könnten. Etwa um fehlende Arbeitskräfte zu kompensieren, indem wir die bestehenden Arbeitskräfte ein bisschen effektiver machen. Ich glaube, da kann man sehr viel erreichen, wenn man einmal anfängt, in diese Richtung zu gucken. Aber solange der Blick starr auf Superintelligenz gerichtet ist, werden viele dieser Richtungen gar nicht ernsthaft verfolgt und das ist schade. Und ich glaube, das ist auch eine verpasste Marktchance. Eine Marktchance, bei der wir in Europa durchaus punkten können. Und zwar, indem wir uns für eine wirtschaftliche Realität nach dieser Skalierungsphase positionieren, in der man glaubt, mit immer mehr kriegt man Superintelligenz, die dafür sorgt, dass wir alle nur noch am Strand liegen und Cocktails trinken.

Moderatorin [00:49:26]

Zum Ressourcenbedarf habe ich tatsächlich noch eine sehr passende Frage von den Journalist:innen. Vielleicht können Sie das relativ schnell beantworten: Wie elegant ist denn die Software, die von KI heute erstellt wird? Könnte da häufig Code entstehen, der zwar funktioniert, aber in „aufgebläht" ist? Und was bedeutet das für den Ressourcenbedarf?

Holger Hoos [00:49:45]

Genau so ist es und das bedeutet natürlich nichts Gutes. Zum einen der Ressourcenbedarf dieser Software, den sie braucht, um einfach nur zu laufen. Wir reden ja zum Teil von Software, die Millionen Mal pro Sekunde aufgerufen wird für Websuchen zum Beispiel. Oder über Systeme, die mit ganz vielen Daten von ganz vielen Bürgerinnen und Bürgern umgehen müssen. Und dann ist es natürlich auch so: Der Aufwand, den man in die Wartung und Erweiterung solcher Systeme stecken muss, der ist auch ganz erheblich, selbst wenn man das durch KI-Systeme machen lässt. Und da glaube ich, wird man demnächst in eine Situation kommen, in der man sieht, dass menschliche Entwicklerinnen und Entwickler im Bereich Software eben doch noch eine ganze Menge Fähigkeiten haben, die KI alleine nicht erbringen kann. Ich glaube, es ist wie fast überall: Im Zusammenspiel liegt nachher die beste Lösung. Wirklich elegante, effiziente algorithmische Lösungen zu finden, ich glaube, da können KI-Systeme bei helfen, aber da braucht es auch durchaus noch menschliche Expertise. Und ich sehe auch nicht, dass sich das in der absehbaren Zukunft ändert. Also viele haben jetzt all ihre Softwareentwicklerinnen und Entwickler entlassen oder sagen, das muss man nicht mehr studieren. Das ist wie das, was ein berühmter Kollege von mir vor zehn Jahren über die Radiologen gesagt hat: Brauchen wir nicht mehr auszubilden. Und wie ist es heute? Zu wenig Radiologen, ne. Also ich denke, dass die Firmen, die das machen, das ganz sicherlich bereuen werden und dafür sehen wir auch schon erste Ansätze.

Moderatorin [00:51:08]

Wir sind jetzt kurz vor Ende. Ich hab aber trotzdem für Sie beide noch eine Frage, die so ein bisschen den Blick in die Zukunft schafft. Und zwar gerne erst Sie, Frau Simon. Wenn Sie sich etwas wünschen könnten von der Politik in Bezug auf KI-Regulierung, was wäre das?

Judith Simon [00:51:27]

Boah, das ist eine schwierige Frage. Ich glaube, tatsächlich, was ich mir wünschen würde, wäre genau diese Haftung. Und zwar diese starke Haftung von Personen. Und ich weiß, ich hab das neulich schon mal gesagt und vielleicht hab ich es nicht bis zum Ende durchdacht, aber meine Überlegung war: Wenn die Leute, die im Moment das Geld mit diesen Unternehmen machen, Sorge haben müssten, dass ihnen nicht nur Geld weggenommen wird, sondern sie ins Gefängnis müssten und es ihre Lebenszeit kostet. Das scheint mir der einzige Hebel zu sein, damit Leute über die Konsequenzen ihrer Technologien nachdenken. Ist es realistisch? Ich weiß es nicht, aber ich habe das Gefühl, dass, wenn man nur über Geld nachdenkt, im Moment diese Akteure teilweise so viel Geld haben, dass selbst das als Hebel nicht funktioniert. Aber diese Sorge, vielleicht wirklich verurteilt zu werden und massive Schäden zu erleiden, würde dazu führen, dass diese Sorgfalt in diese Systeme Einzug erhält und auch vielleicht diese Beschleunigung gebremst wird. Denn im Moment sagen auch beim Digitalgipfel die Leute von Anthropic selbst, sie kommen nicht hinterher mit ihren Safety Measures, weil die Entwicklung so schnell ist. Diese Entschleunigung nicht nur zu fordern, sondern, dass sie tatsächlich passiert, hängt, glaube ich, an dieser persönlichen Haftung.

Moderatorin [00:52:35]

Und vielen Dank. Und die Frage an Sie, Herr Hoos, wenn Sie sich so die Weiterentwicklung von KI-Modellen angucken, in welche Richtung sollte das gehen? Sollte man wirklich versuchen, die großen Alleskönner-Modelle immer besser zu machen, auf die Gefahr hin, dass man sie dann eben nicht kontrollieren kann? Oder ist die Zukunft eher in Richtung kleinere, spezialisierte Modelle aus Ihrer Sicht?

Holger Hoos [00:52:56]

Also ich denke, die letztere Richtung ist auf jeden Fall eine, die man verfolgen sollte und auch mehr, als das zurzeit der Fall ist. Denn da ist auf eine ungefährlichere, und ich sag jetzt nicht nur ungefährlicher in diesem philosophischen Sinn, sondern auch einfach wirtschaftlich, haftungstechnisch, schadentechnisch ungefährlichere Art und Weise wirklich guter Nutzen zu erzielen. Was ich mir wünschen würde, ist, dass es eine bessere Sicherheitskultur in der Ingenieursdisziplin gibt, die uns diese KI-Systeme beschert. Ich find es einfach unfassbar, dass eine derart mächtige Technologie eigentlich ohne eine gute Sicherheitskultur entwickelt wird und auch ohne das Bewusstsein, dass man so eine Sicherheitskultur braucht. In anderen Bereichen ist das anders, in denen es sagen wir mal, wenn die Technologie einem entgleist, große Schäden gibt. Ein viel zitiertes Beispiel ist natürlich der ganze Nuklearbereich, der jetzt bei uns in Deutschland nicht mehr so diskutiert wird. Aber wir haben ja jenseits der Grenzen jede Menge Kernkraftwerke. Da gibt es eine sehr gute Sicherheitskultur. Die ist immer noch nicht perfekt, weshalb viele Leute sich vielleicht begreiflicherweise immer noch unwohl damit fühlen. Aber es gibt eine gut ausgeprägte Sicherheitskultur und im Bereich KI gibt es die nur sehr eingeschränkt. Und ich denke, die brauchen wir ganz dringend. Ich würde mir sehr wünschen, dass die nicht erst dann entsteht, wenn wir die ersten großen Super-GAUs erlebt haben und die ersten großen Unfälle, viele Schäden verursacht haben. Sondern, dass das vielleicht ein bisschen vorher passiert. Und obwohl ich grundsätzlicher

Weise ein Optimist in dieser Frage bin, fürchte ich, werden wir uns eine noch wesentlich blutigere Nase holen, bevor diese Sicherheitskultur tatsächlich entsteht. Und das finde ich bedauerlich.

Moderatorin [00:54:37]

Alles klar. Ja, vielen Dank Ihnen beiden auch noch mal. Obwohl er nicht mehr da ist: vielen Dank an Herrn Geiping. Hoffentlich sieht er das noch mal. Vielen Dank auch an die Journalistinnen und Journalisten für Ihre Aufmerksamkeit und für Ihre Fragen. Wie ich am Anfang erwähnt habe, stellen wir eine Audio-, Videodatei und ein maschinell erstelltes Transkript so schnell wie möglich jetzt nach Ende dieses Meetings zur Verfügung. Im Laufe des Tages gibt es dann noch ein redigiertes Transkript. Sie finden das alles über den Link in der Einladungsmail von gestern. Und dann kann ich nur noch sagen: einen schönen Tag noch. Ich hoffe, wir sind alle gut in den Tag gestartet so früh und bis zum nächsten Mal. Tschüss.

Holger Hoos [00:55:12]

Tschüss.

Ansprechpartnerin in der Redaktion

Samantha Hofmann
Redakteurin für Digitales und Technologie

Telefon +49 221 8888 25-0
E-Mail redaktion@sciencemediacenter.de

Impressum

Die Science Media Center Germany gGmbH (SMC) liefert Journalisten schnellen Zugang zu Stellungnahmen und Bewertungen von Experten aus der Wissenschaft – vor allem dann, wenn neuartige, ambivalente oder umstrittene Erkenntnisse aus der Wissenschaft Schlagzeilen machen oder wissenschaftliches Wissen helfen kann, aktuelle Ereignisse einzuordnen. Die Gründung geht auf eine Initiative der Wissenschafts-Pressekonferenz e.V. zurück und wurde möglich durch eine Förderzusage der Klaus Tschira Stiftung gGmbH.

Nähere Informationen: **www.sciencemediacenter.de**

Diensteanbieter im Sinne MStV/TMG
Science Media Center Germany gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
Amtsgericht Mannheim
HRB 335493

Redaktionssitz
Science Media Center Germany gGmbH
Rosenstr. 42–44
50678 Köln

Vertretungsberechtigter Geschäftsführer
Volker Stollorz

Verantwortlich für das redaktionelle Angebot (Webmaster) im Sinne des §18 Abs.2 MStV
Volker Stollorz

science
media center
germany