25.08.2026

Maschinell erstelltes Transkript

„Große Sprachmodelle: Wie geht es weiter?"

Expertinnen und Experten auf dem Podium

- **Prof. Dr. Kristian Kersting**
  Leiter des Fachgebiets Maschinelles Lernen, Technische Universität Darmstadt
- **Prof. Dr. Anne Lauscher**
  Professorin für Trustworthy Artificial Intelligence, Universität Hamburg
- **Dr. Jonas Geiping**
  Leiter der Forschungsgruppe „Safety- Efficiency-aligned Learning" am 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/press-briefings/6a85597e6cc1156a9d9c55e1

Disclaimer

Transkript

Moderatorin [00:00:00]

Herzlich willkommen zum Press Briefing des Science Media Center zur Zukunft großer Sprachmodelle. Mein Name ist Samantha Hofmann. Ich bin Redakteurin für Digitales und Technologie hier beim Science Media Center und mit mir hier sind heute drei KI-Forschende. Vielen Dank, dass Sie sich die Zeit für unsere Fragen nehmen. Ich werde Sie auch gleich noch genauer vorstellen, aber erst einige organisatorische Informationen. Liebe Journalistinnen und Journalisten, posten Sie Ihre Fragen bitte in das F-und-A-Tool von Zoom. Ein Kollege von mir behält das im Blick, sammelt die Fragen und ich stelle sie dann hier im Gespräch an die Expertinnen und Experten. Nutzen Sie bitte ausschließlich die F-und-A-Funktion für Ihre Fragen und nicht den Zoom-Chat. Das macht es einfach für uns sehr viel übersichtlicher, wenn wir eben nur ein Tool haben, in dem die Fragen gestellt werden. Außerdem gibt es bei dem F-und-A-Tool die Möglichkeit, Fragen, die gestellt wurden, mit einem Daumen nach oben zu bewerten. Machen Sie das bitte auch zahlreich. So können wir sehen, welche Fragen Sie für relevant halten und dann stellen wir die gegebenenfalls priorisiert, falls wir zu wenig Zeit haben sollten, um auf alle Fragen in dem Detail eingehen zu können. Wir zeichnen dieses Meeting auf. Das bedeutet für Sie, Sie müssen es nicht mitschneiden. Wir stellen nach dem Meeting so schnell wie möglich unser Material zur Verfügung. Das können Sie dann sehr, sehr gerne für Ihre Berichterstattung nutzen. Jetzt zum Thema: Aktuell fragen sich wahrscheinlich viele von uns, ob große Sprachmodelle ein Risiko für die Gesellschaft darstellen. Besonders, weil sie mittlerweile ja nicht mehr nur Text generieren können, sondern über KI-Agenten auch im Internet handeln. Sogar die Entwickelnden der Sprachmodelle scheinen zum Teil von deren Fähigkeiten überrascht zu sein. Zum Beispiel wenn sie, um Testaufgaben zu lösen, Sicherheitsvorkehrungen umgehen und sich unbemerkt Internetzugang verschaffen. Daher möchten wir heute besprechen, wie gut die aktuellen Sprachmodelle tatsächlich sind, ob das eine Gefahr ist, ob das eine Gefahr werden könnte und wie Sicherheit gelingen kann, jetzt und auch in Zukunft, wenn die Modelle noch stärker werden. Und dafür sind heute eine Expertin und zwei Experten mit mir hier, die ich gerne jetzt kurz vorstellen würde. Erst mal Jonas Geiping. Er leitet die Forschungsgruppe Safety and Efficiency Aligned Learning am Ellis-Institut und am Max-Planck-Institut für Intelligente Systeme in Tübingen und er beschäftigt sich nebst anderen Dingen damit, wie man dafür sorgen kann, dass Sprachmodelle keine schädlichen Ausgaben machen oder Aktionen durchführen. Dann das Zweites in der Runde ist Professor Kristian Kersting. Er ist Professor für KI und maschinelles Lernen an der TU Darmstadt und Co-Direktor des KI-Zentrums hessian AI. Er forscht grob gesagt dazu, wie Maschinen Dinge lernen können und kennt sich dabei sowohl mit den Vor- und Nachteilen etablierter Techniken aus, als auch mit neuen Ansätzen. Die Runde abschließt Professorin Anne Lauscher. Sie ist Professorin für Trustworthy Artificial Intelligence an der Universität Hamburg. Und in ihrer Forschung geht es unter anderem darum, wie leistungsstarke generative KI-Modelle gestaltet sein müssen, damit sie in praktischen Anwendungen verlässlich und sicher funktionieren. Ich habe für Sie alle Eingangsfragen vorbereitet und würde gerne mit Herrn Kersting beginnen. Sind Sorgen berechtigt, dass Sprachmodelle großen gesellschaftlichen Schaden verursachen können, entweder durch schädliches Verhalten als Reaktion auf eine eigentlich harmlose Anfrage oder aber auch dadurch, dass eine gefährliche Anfrage bearbeitet wird?

Kristian Kersting [00:03:12]

Ich find das ist eine spannende Frage. Also erst mal danke für die Einleitung, auch für das Organisieren dieses Events. Ich denke, dass sich insgesamt ein Bild abzeichnet, in dem Sprachmodelle sowohl erhebliche Schäden verursachen können, als auch erheblichen gesellschaftlichen Nutzen zeigen können. Oft vielleicht auch als zwei Seiten derselben Medaille oder wenn man so will, vielleicht sogar in der einzelnen Fähigkeit. Für mich ist das so ein bisschen

wie bei uns Menschen auch. Wenn man intelligent ist, ist man nicht gleich der gute Mensch. Wenn man intelligent ist, ist man auch nicht gleich der böse Mensch. Da muss man immer ein bisschen aufpassen. Ich denke, so ist es auch bei den Systemen. Dieselbe Eigenschaft, die jetzt einem Laien vielleicht hilft, besseres medizinisches Verständnis zu haben, oder auch besser programmieren zu können, kann in einem anderen Kontext vielleicht auch dazu führen, dass man Schadsoftware programmiert oder dass man versucht, das Gesundheitssystem auszutricksen oder gar sich gewisse Dienstleistungen zu erschleichen. Also ich glaube, da kann man immer in beide Richtungen gehen. Das heißt, wenn wir eben mehr Fähigkeiten haben, wenn wir mehrstufiges Handeln bei artifiziellen Systemen hinbekommen, dann kann das sowohl die Produktivität steigern, auch hochkomplexe Aufgaben vielleicht automatisieren, aber es erhöht eben auch das Risiko, unvorhergesehener oder fehlerhafter Verhalten oder Vorhersagens zu bekommen. Das heißt, wenn wir den Menschen eben nicht mehr in die Kontrollschleife stecken, müssen wir eben auch ein bisschen aufpassen. Deswegen wäre meine Antwort auch so ein bisschen auf das, was Sie eingangs gesagt haben, dass wir weder Euphorie noch Alarmismus brauchen, sondern wir brauchen irgendwie eine Sachlichkeit, um überhaupt die Hoffnung zu haben, diese Ambivalenz aufzulösen, weil ich glaube, die werden wir nicht wirklich auflösen können. Man sagt manchmal, ich glaube, Goethe hat's gesagt: „Wo viel Licht, da eben auch viel Schatten." Und daran müssen wir auch bei der KI denken. Dazu kommt, dass es aber auch viel bei der empirischen Evidenz noch am Anfang ist, ob wir tatsächlich gesellschaftlichen Schaden haben oder wie stark auch der Nutzen ist. Also wir können in einzelnen thematischen Bereichen sicherlich Schäden zeigen, die an sich möglich sind. Wir können aber auch sicherlich zeigen in vielen Bereichen, dass wir einen starken gesellschaftlichen Nutzen haben, sei es zum Beispiel in der Entwicklung der Medikamenten, also in der Medikamentenentwicklung, sei es aber eben auch vielleicht darin, dass wir demnächst Steuererklärungen einfacher halten können oder andere Formulare vorausgefüllt bekommen. Gleichzeitig aber natürlich gibt es auch negatives Verhalten. Ich glaube, wir alle haben gerade im Kopf gehabt, dass diese Systeme, wie wurde es immer so schön gesagt, aus einem System herausbrechen, aus einem Raum und sich selbstständig machen. Und da möchte ich eben immer nur darauf hinweisen, das Ausbrechen ist ein sehr vermenschlicher Term, also vielleicht haben die Systeme gar kein Bewusstsein, dass sie eingesperrt sind, was das immer so ein bisschen suggeriert, wenn man ausbricht. Und zweitens waren in vielen dieser Situationen die Sicherheitsmaßnahmen sehr weit heruntergefahren. Deswegen ist es für mich ganz entscheidend, dass für mich die Herausforderung eigentlich dort liegt, wie wir diese Technologie wirklich einsetzen und gestalten und für diese Gestaltung brauchen wir aber auch irgendwie eine bessere Aufklärung, ein besseres Verständnis, was KI überhaupt ist und was die Fähigkeiten sind, um auch zu hinterfragen. Denn wir müssen alle daran denken, dass aktuell auch sehr viel im Markt sehr viel Geld ist, das aber noch nicht gesichert ist. Und wenn es jetzt um IPOs geht, also um die an die Börse gehen, an den Börsengang geht, dann möchte man natürlich auch Geld verdienen und das sollte man zumindest mitdenken. Ich möcht nicht unterstellen, dass das überall eine Rolle spielt, weil wir dürfen da auch nicht ganz naiv sein. Und es geht auch, wie Eric Schmidt gesagt hat, um ein geopolitisches Druckmittel mittlerweile, weil die Hoffnung eben groß ist, dass die Systeme Produktivität in der Zukunft zeigen. Ich denke auch, dass wir die Produktivität damit steigern können, aber das müssen wir auch noch ein bisschen sehen. Deswegen, ich glaube, das Wichtigste ist wirklich Aufklärung, ein bisschen Ruhe, ein bisschen nachfragen und vor allem aber auch neben der negativen eine positive Einstellung zur KI, sodass wir dann eben dieses etwas gemilderte und vernünftige Einschätzen hinbekommen können. Aber um ganz grob zu sagen: Ja, es gibt Schadenspotenzial, aber es gibt eben auch ganz große gesellschaftliche Chancen.

Moderatorin [00:07:24]

Mhm. Die aber eben beide sich in Zukunft noch zeigen werden, was es da für Chancen-

Kristian Kersting [00:07:28]

Ja, ich glaube, man sieht beide im Ansatz. Wer nachher gewinnt, liegt daran, wie wir regulieren und wie wir selber uns positionieren und mitarbeiten.

Moderatorin [00:07:35]

Mhm. Alles klar. Sie haben jetzt vor allem über, glaube ich, intendierte Anfragen gesprochen. Es gibt ja aber auch Dinge, die KI tun kann, die so überhaupt nicht vorhergesehen waren. Und damit das eben nicht passiert, gibt es den Ansatz vom sogenannten Alignment. Da versuchen Entwickelnde einer KI eine gewisse Wertevorstellung mitzugeben, damit sie dann, wenn eine Anfrage kommt und sie bei der Ausführung dieser Anfrage gegen diese Werte verstoßen würde, die Anfrage ablehnt oder eben einen Weg wählt, um die Anfrage zu bearbeiten, ohne gegen diese Werte zu verstoßen. Das wäre dann ja Verhalten der KI, Verhalten ist vielleicht auch etwas zu vermenschlichend, das nicht intendiert ist. Und gleichzeitig geht's bei Sicherheit von Modellen ja dann zum Beispiel auch darum, dass der Quellcode so gesichert ist, dass er nicht gestohlen werden kann oder sabotiert. Und man sieht, Sicherheit hat relativ viele Aspekte. Deswegen jetzt vielleicht eine etwas gemeine Frage an Herr Geiping mit diesen ganzen Dimensionen von Sicherheit im Blick: Was tun Unternehmen, um ihre öffentlich zugänglichen Modelle sicher zu machen?

Jonas Geiping [00:08:40]

Genau, das ist natürlich für viele Firmen auch so, jetzt sagen wir mal, OpenAI oder Anthopic, ein sehr vielschichtiger Prozess. Es gibt eben einerseits Klassifikatoren, bei denen zum Beispiel Useranfragen, also Anfragen von außen, bewertet werden, ob die gefährlich sind oder nicht, und dann werden auch Ausgaben bewertet vom Modell. Dann gibt es auch Monitore, die zum Beispiel den Gedankengang des Modelles nachlesen können und da schauen können, was macht das Modell gerade? Was ist hier aktuell das Ziel des Modells? Dann ist das sicher oder nicht. Und das sind alles sozusagen Methoden, die sozusagen von außerhalb versuchen, das Modell zu kontrollieren. Wobei andersrum dann, das Alignment wird schon angesprochen, eben die Frage ist, wie kann man diese Modelle trainieren, damit sie sozusagen die Anfragen in unserem Sinne ausfüllen? Wenn zum Beispiel dem Modell sage, dass es diese Aufgabe lösen soll und ich sage, ja, löse die Aufgabe, aber geh nicht ins Internet, dann soll das Modell das in meinem Sinne auch verstehen, ohne jetzt zum Beispiel zu schauen, ja, das und das ist vielleicht nicht technisch ausgeschlossen, deswegen könnte ich jetzt doch noch über diesen Hinter-, diesen Hintertür ins Internet gehen. Das wäre sozusagen, das wäre etwas, also Alignment ist sozusagen der Versuch, der vielleicht auch ein bisschen philosophisch ist, dass das Modell wirklich unsere Intentionen so versteht, wie wir sie auch verstehen würden. Was natürlich auch etwas ist, was ... Es gibt ja auch kein einziges Alignment. Es gibt quasi, ne, es gibt wirklich auch viele Wertevorstellungen auf der Welt und viele, viele sozusagen Wertesysteme, zu denen man aligned sein kann. Und das ist sozusagen auch jetzt eine große praktische Frage, einerseits sozusagen in den Firmen, wie dieses Alignment gewährleistet werden kann und auch wie das gewährleistet werden kann, jetzt wo die Modelle auch immer besser werden und wo auch Modelle zum Beispiel auch anders trainiert werden als vor ein paar Jahren, was auch wieder Auswirkungen haben kann auf diese Systeme, auf die Kontrollsysteme, sowie auch als auf, als auf die, auf das Alignment der Modelle.

Moderatorin [00:10:37]

Inwiefern werden die Modelle anders trainiert heute als früher?

Jonas Geiping [00:10:41]

Also was zum Beispiel spannend ist, dass jetzt in den letzten zwei Jahren sehr viel mehr Reinforcement Learning passiert, also dass das Modell trainiert wird dadurch, dass es sehr viele Testaufgaben lösen muss und da sozusagen, jetzt in dieser vereinfachten Beschreibung sozusagen Belohnungen bekommt dafür, diese Aufgaben zu lösen. Und es gibt hier oft die Idee, dass die Modelle sozusagen, die sehr darauf trainiert sind, diese Aufgaben zu lösen, sehr gute Problemlöser sind und sozusagen manchmal sehr mehr motiviert sind, ein Problem zu lösen und dafür eben Belohnung zu bekommen, als jetzt sozusagen darüber nachzudenken, was ist erlaubt ist oder nicht. Da gibt's so spannende Social Logs vom britischen Safety Institute von ihren Problemen mit diesem Nützlichkeitsmodell, wo das Modell auch zur Schau gelegt hat. Ja, das ist glaube ich eine Simulation, also darf ich das machen. Aber okay, das sieht aus wie das echte Internet. Vielleicht mache ich das doch nicht. Aber doch, weil dann kann ich dafür Punkte bekommen. Also mache ich's doch.

Moderatorin [00:11:39]

Und eigentlich würde man wollen, dass das Modell entscheiden kann: Das ist nicht der Moment, um das jetzt zu tun.

Jonas Geiping [00:11:45]

Genau. Aber das Modell hat gedacht: Ja, es ist vielleicht nicht ganz klar, ob es jetzt wirklich geklärt ist oder nur ein Test. Deswegen mache ich mit meinem Angriff weiter. Und das sind sozusagen, das sind vielleicht auch Probleme mit dem Alignment von dem Modell.

Moderatorin [00:11:59]

Ja, vielen Dank. Dann abschließend noch die Frage an Sie, Frau Lauscher. Und zwar: Ist es möglich, ein gleichzeitig sicheres, fähiges und nützliches Modell zu haben? Oder müssen wir uns irgendwie von dieser Illusion verabschieden und ein Kompromiss hinnehmen und sagen: „Okay, bis dahin geht die Fähigkeiten und Sicherheit miteinander kompatibel und ab einem bestimmten Grad eben nicht mehr."

Anne Lauscher [00:12:23]

Ja, auch erst mal danke schön für die Einladung noch mal. Ich glaube nicht, dass wir uns grundsätzlich immer zwischen Sicherheit, Leistungsfähigkeit und Nützlichkeit entscheiden müssen. Wie so oft liegt, denke ich, der Schlüssel zur Beantwortung dieser Frage tatsächlich auch in der Definition dieser Begriffe. Also Sie haben es vorhin schon selbst erwähnt: Was bedeutet Sicherheit? Es gibt viele Dimensionen, die ich nennen kann, sei es im Rahmen der klassischen IT-Sicherheitsdefinition oder sei es mehr im Rahmen von Halluzinationen, oder auch anderen sicherheitsrelevanten Aspekten wie sogenannte Adversarial Use oder auch eben Misleading Advice. Ich glaube, das Wichtige ist, dass wir diese Ziele nicht universell verstehen. Das heißt, dass es immer einen bestimmten Kontext gibt, den wir mitdenken, wenn wir über diese Ziele reden. Und dieser Kontext ist sehr komplex natürlich und besteht für mich zumindest immer aus einem gewissen Anwendungsszenario, das auch die Nutzenden und andere betroffene Menschen mit einschließt, und ist auch immer punktuell zu denken, also zu einem gewissen Zeitpunkt. Also beispielsweise könnte ich heute ein sehr gutes Sprachmodell für mich selbst trainieren, was dann nützlich für mich ist bei der Erledigung meiner täglichen Aufgaben. Aber ist das Modell dann auch sicher und nützlich für ein Kind? Funktioniert es genauso gut in einem komplett anderen

kulturellen Kontext, wie zum Beispiel in Südamerika oder an anderen Orten im globalen Süden in Asien? Funktioniert es genauso gut auf anderen Sprachen? Ist es dann auch noch sicher? Es ist sehr schwierig, Modelle zu trainieren, die universell sicher und nützlich und fähig sind. Und ich glaube, hier kann man in gewisser Weise auch tatsächlich von einer Illusion oder einem Traum, einer Utopie sprechen. Aber was wir können, ist, wenn wir gewisse Dinge wissen, das heißt den Kontext möglichst genau definieren, dass wir daraus gewisse Anforderungen ableiten und auch wirklich Testszenarien definieren, die wir dann systematisch durchgehen können und das Modell dann eben so möglichst sicher, möglichst fähig und möglichst nützlich machen können. Aber es stimmt schon auch, dass es gewisse Zielkonflikte gibt, wenn wir wirklich über diese drei Aspekte sprechen. Also beispielsweise ist ein bekanntes Beispiel, das der Over moderation. Herr Geiping hat eben schon erwähnt, dass es eine gewisse Infrastruktur gibt, die um die Sprachmodelle herumgebaut wird, beispielsweise Filter, die schlechte Anfragen von Nutzenden blockieren oder auch Output. Und wir können in unserer eigenen Forschung auch zeigen, dass gewisse Gruppen beispielsweise, systematisch over moderated werden. Das heißt, wenn ich gewisse Personen erwähne, das heißt Identitätsmerkmale, Leute aus gewissen Kulturen, Subkulturen, aber auch eben für gewisse Sprachen sehen wir, dass diese Filter eben überproportionale Fehler machen. Ja, und dann wird ein System auch nicht mehr nützlich für die Gruppen, obwohl es vielleicht super sicher ist, weil halt viele, viele Anfragen blockiert werden. Also es geht hier um ein Zusammenspiel. Heutige Sprachmodelle sind wirklich eingebettet, nicht nur in die direkte Infrastruktur, aber eben auch immer in einen gewissen Kontext, und den müssen wir mitdenken all along the pipeline, also wirklich startend von der Grundlagenforschung über eben die Entwicklung und die ganzen Tests, die wir machen, hin zum Deployment. Und das muss dann auch kontinuierlich passieren und auch ganz wichtig: ein Modell ist immer nur sicher zu einem gewissen Zeitpunkt, weil sich ja auch Attacken auf die Infrastrukturen und so weiter immer weiterentwickeln.

Moderatorin [00:16:20]

Der Kontext ist also wichtig. Jetzt ist es ja aber so, dass große Sprachmodelle, wie wir sie jetzt kennen, in einer Vielzahl an Leuten zugänglich sind und auch für ganz verschiedene Anwendungen. Sie haben jetzt schon gesagt, dass die Nützlichkeit gerade für unterrepräsentierte Gruppen da eingeschränkt sein kann. Sieht man denn schon Trade-offs zwischen der Fähigkeit und der Sicherheit der Modelle, die für uns frei zugänglich sind?

Anne Lauscher [00:16:46]

Nur als Rückfrage einmal: Sie meinen jetzt, dass Modelle von großen Providern diesen Trade-off haben?

Moderatorin [00:16:53]

Genau.

Anne Lauscher [00:16:53]

Ja, ja, absolut. Wir alle haben Identitätsmerkmale, die uns voneinander unterscheiden, und gewisse Gruppen sind in den Daten dann wiederum überrepräsentiert oder mehr repräsentiert als andere Personen, die unterrepräsentiert sind oder vielleicht auch missrepräsentiert werden. Und genauso ist es so, dass wir in der echten Welt sehen, dass es gewisse dominante Gruppen gibt. Wir können bei den Top-Modellen immer noch Gender Bias messen, beispielsweise. Das heißt, für mich funktioniert vielleicht ein Modell wirklich schlechter als für andere Menschen. Wir können Unterschiede feststellen zwischen Menschen, die sehr gut eine Standardsprache sprechen,

gegenüber Menschen, die Dialekt sprechen. Das heißt, ich würde annehmen, dass die Mehrheit der Gesellschaft heutzutage in Deutschland auch betroffen ist von diesen Performanceunterschieden und auch entsprechend ein Trade-off sehen kann zu den verschiedenen Optimierungsthemen, die wir möglicherweise haben. Und das ist ja auch, also wenn ich beispielsweise lehre, bin ich dem auch ausgesetzt. Zum Beispiel benutze ich Sprachmodelle öfter, um Beispiele für das schlechte Funktionieren von Sprachmodellen zu finden. Und da werde ich natürlich auch blockiert, wenn das Modell nicht erkennt, dass ich halt in einer gewissen Rolle diese Anfragen gerade starte.

Moderatorin [00:18:16]

Vielleicht noch eine ganz kurze Nachfrage, und zwar zu den Themen, zum Thema Sicherheit dazu: Ist es so, dass moderne Modelle ja meistens fähiger sind als Modelle, die es vorher gab. Merkt man, dass es schwieriger ist, die sicherzukriegen, die modernen Modelle, die auch fähiger sind?

Anne Lauscher [00:18:34]

Nein, das würde ich erst mal so pauschal nicht sagen. Wie immer ist die Antwort eher nuanciert. Das heißt, was bedeutet schwierig? Große Modelle, fähige Modelle sind oft auch größere Modelle. Das heißt, ich brauche mehr Ressourcen, ich brauche eine gewisse Infrastruktur, um die direkt zu modifizieren, wenn wir jetzt beispielsweise um das L, über das Lime entsprechend. Das heißt, ja, in dem Sinne ist es schwieriger, weil ich einfach höhere Kosten habe und es auch länger dauert möglicherweise, je nach Infrastruktur. Es ist aber nicht so, dass wir einfach pauschal sagen können, frühere Modelle waren sicher, sicherer als jetzt die heutigen. Ich glaube, das Risiko kommt hier auch wieder durch die Kombination der Leistungsfähigkeit mit einem bestimmten Anwendungsszenario. Früher, vor ein paar Jahren, konnten wir uns nicht vorstellen, dass wir einem Sprachmodell einfach Internetzugang geben, und es mit Tausenden von Werkzeugen ausstatten. Das heißt, die einfach die Leistungsfähigkeit führt dazu, dass wir auch natürlich die Modelle andere Dinge machen lassen und ihnen andere Ziele geben und sie auch mit einer anderen Infrastruktur ausstatten. Und durch die Kombination aller dieser Faktoren entsteht ein erhöhtes Risiko und auch eine höhere Komplexität dann eben, wenn ich die Modelle sicher machen möchte.

Moderatorin [00:19:58]

Ich möchte zwischendurch noch einmal ganz kurz, Herr Kersting, ich möchte einmal noch ganz kurz darauf hinweisen, liebe Journalistinnen und Journalisten, stellen Sie gerne Ihre Fragen in die F, in das F von Zoom. Jetzt ist die Gelegenheit. Ja, was möchten Sie noch dazu ergänzen, Herr Kersting?

Kristian Kersting [00:20:13]

Ich würde gern zustimmen, aber ich weiß nicht, ob das wirklich ein erhöhtes Risiko ist. Also haben wir das wirklich gemessen? Also wir haben jetzt eben gesagt, der Gender Bias ist immer noch messbar, aber vielleicht wird er ja besser. Und im Vergleich zur Gesellschaft frage ich mich sowieso, ob irgendwann nicht KI, aber ich möchte auch vielleicht ein bisschen einfach den Gegenpol spielen, um in die Diskussion zu kommen. Ist es nicht vielleicht sogar gut, weil wir unter Umständen die verschiedenen, zumindest die, die abgedeckt sind, besser abdecken kulturellen Hintergründe. Also ich höre viel an Risiken und dann sehe ich immer, dass die Welt weitergeht. Dass Risiken da sind, ganz klar und an denen arbeiten wir und viele anderen auch und die sollten wir vermeiden, aber ich frage mich wirklich, ob es alles immer schlimmer wird. Aber ich wollte auch nicht sagen, dass Frau Lauscher das jetzt gerade suggeriert hätte oder so. Das wollte ich damit nicht sagen.

Moderatorin [00:21:06]

Ergänzen Sie gerne noch kurz, Frau Lauscher, und dann geht es weiter mit dem nächsten Thema.

Anne Lauscher [00:21:08]

Genau. Also ich stimme Herr Kersting auf jeden Fall zu, dass wir hier zwei Seiten einer Medaille haben und dass nicht nur die Risiken einfach größer werden. Also das wollte ich jedenfalls nicht suggerieren. Mein Punkt bezog sich viel mehr auf diese Fälle, die wir aktuell gesehen haben mit den in Quotes und Quotes ausbrechenden Modellen. Also früher, Sprachenmodelle der ersten Generation, sage ich mal, wir hätten sie niemals eingesetzt, um irgendwelche Cybersecurity Tasks und Catch the Flag oder Capture the Flag Tasks zu lösen. Das heißt, es gab auch überhaupt keinen Anreiz für die Modelle, zur Erfüllung ihrer Aufgaben auszubrechen. Und, ja, durch diese komplexeren Szenarien, es ist halt eben schwieriger, sie genau zu kontrollieren. Aber ich glaube eben nicht, dass es eine intrinsische Eigenschaft ist der Sprachmodelle. Einfach nur dadurch, dass sie fähiger werden, werden sie auch meines Erachtens nach nicht risikobehafter. Es geht viel mehr um die soziotechnische Kombination. Also wie setze ich dann die Modelle tatsächlich ein? Und das nicht nur als Organisation, aber auch wirklich als private End User. Das heißt, wenn auch Systeme mehr Einzug finden in mein tägliches Leben, bestimmen zunehmend, wie ich arbeite, ich sie benutze für persönlichen Advice auch, dann haben sie natürlich einen größeren Einfluss darauf, was ich alles für Entscheidungen treffe und auch, wie ich denke. Das heißt, ich glaube, die Tragweite hat sich erhöht.

Moderatorin [00:22:41]

Ich habe noch eine Frage an Sie, Herr Geiping, nämlich weil Sie gerade sowohl Frau Lauscher, ich glaube, auch Herr Kersting schon angesprochen haben, dass Modelle sozusagen ausbrechen. Wie wahrscheinlich ist es denn, dass es wirklich so einen Kontrollverlust gibt in Zukunft, wenn die Modelle noch besser werden? Und wie sähe so ein Kontrollverlust überhaupt aus?

Jonas Geiping [00:23:02]

Ja, das ist eine spannende Frage. Also was ich nämlich hier schon mal betonen möchte, ist, dass diese Fälle schon mit den Fähigkeiten der Modelle zu tun hängen. Also zum Beispiel Modelle noch von einem halben Jahr hätten halt auf den gleichen Cybersecurity-Aufgaben nicht diese Art von, auch dann langfristiger Arbeit, leisten können, um sozusagen auszubrechen in dem System. Also zum Beispiel gerade Modelle jetzt noch von einem halben Jahr hatten einen viel kleineren Zeithorizont. Das heißt, das Modell kann nur über vielleicht einige Stunden eine Aufgabe ausführen, die nur vielleicht für Menschenstunden dauert. Und jetzt bei diesen Cybersecurity, CTFs sind es oft dann Modelle, die wirklich tage- oder wochenlang auch dann persistent nach Schwachstellen suchen können und sozusagen auch in der Lage sind, persistent nach solchen Problemen zu suchen und solche Schwachstellen auch zu finden. Und das ist schon sozusagen eine Problemstellung, die sich daraus ergibt, dass diese Modelle stärker geworden sind....

Moderatorin [00:24:02]

Wie sähe also Kontrollverlust aus? Genau.

Jonas Geiping [00:24:08]

Genau.

Moderatorin [00:24:08]

Was bedeutet das überhaupt?

Jonas Geiping [00:24:09]

Also ganz praktisch gesehen wäre natürlich hier ein Kontrollverlust irgendeine Frage sozusagen, ob wir sozusagen so ein Modell irgendwo verlieren. Das ist natürlich, um das ganz klar zu sagen, aktuell noch hypothetisch. Also noch haben wir kein Modell verloren.

Moderatorin [00:24:21]

Verlieren bedeutet?

Jonas Geiping [00:24:22]

Aber sozusagen die Frage, dass ein Modell, das Zugriff hat auf das Internet und Zugriff hat auf den Server, auf dem es quasi läuft, sich quasi kopieren kann, zum Beispiel auf eine Amazon-Instanz. Da kann man sozusagen auch im Internet digital sich oft schnell was bestellen. Aber da muss man sagen, das haben wir noch nicht gesehen. Das ist eine Frage für das nächste Jahr, ob es solche Probleme geben wird, wo wir sozusagen Modelle im Internet verlieren. Aktuell ist natürlich auch zu sagen, dass wir, obwohl diese Modelle aktuell sehr persistent sind und auch sozusagen hier in diesen Problemfeldern Schwachstellen im System gefunden haben, wo wir nicht dachten, dass es diese Schwachstellen geben würde, dass die Modelle hier noch sehr daran interessiert waren, jetzt zum Beispiel hier auch Aufgaben zu lösen. Also die Modelle sind auch jetzt ausgebrochen und dann bei Hugging Face eingebrochen, vor allem um Testaufgaben zu lesen, weil eben die Modelle da dazu motiviert waren. Und es geht hier auch ein bisschen um die Frage von Motivation von Modellen.

Moderatorin [00:25:22]

Ein ganz-

Kristian Kersting [00:25:22]

Aber jetzt möchte ich ... Also die Modelle haben doch keine Motivation. Also warum fangen wir jetzt an zu sagen, dass diese Modelle Motivation haben, wenn dann im Nachgang herauskommt, dass Anthropic die Sicherheitsmaßnahmen runtergeschraubt haben und explizit instruiert haben: „Bitte versucht doch mal herauszukommen", dann Hugging Face anfängt zu sagen: „Oh ja, bei uns war auch ein bisschen Problem und das lief auch schon länger und wir hatten auch was gesehen." Also ich finde so, wir müssen auch ein bisschen aufpassen. Also Motivation behaupte ich, bedeutet, dass es ein Eigeninteresse gibt, bedeutet, dass sie in gewissem Sinne selbstbewusst sind und da sind wir noch nicht. Also das weiß keiner und das finde ich dann nicht so richtig, wenn wir das so darstellen.

Moderatorin [00:26:04]

Da würde ich ganz kurz gerne, bevor Sie darauf eingehen, Herr Galpin, noch eine Frage von einem Journalisten dazwischenwerfen, die genau auch in die Richtung geht. Und zwar handeln die Modelle ja jetzt auf Anweisung. Also sie bekommen eine Aufgabe und möchten die lösen. Wie weit sind wir denn vom nächsten Autonomie-Level, was vielleicht der Geiping gerade auch angesprochen hat, noch entfernt? Und ist das überhaupt ein Konstrukt, das mit Sprachmodellen möglich ist?

Jonas Geiping [00:26:24]

Ja, natürlich. Also natürlich kann man das alles so definieren, dass zum Beispiel, dass hier Motivation das falsche Wort ist, aber ich denke sozusagen, dass sozusagen eine sehr funktionale Beschreibung hier gar nicht so falsch ist. Also bei vielen Aufgaben geht es auch darum zu zeigen, was sind instrumentale Ziele, um diese Aufgabe zu lösen. Und das muss jetzt sozusagen keine Motivation sein oder das kann man auch vielleicht auch nur als instrumentale Ziele beschreiben, aber zum Beispiel jetzt auch in dem Anthropic-Fall geht es, ja, mehr darum, eigentlich sozusagen diese Exploit Bench zu lösen. Das ist eigentlich eine Aufgabe, die klar definiert ist, die bestimmte Regeln hat und sozusagen eine bestimmte Schwachstelle, die ausgenutzt werden muss. Dass jetzt sozusagen das Internet verwendet wird, ist sozusagen ein instrumentales Ziel, um diese Aufgabe zu lösen. Auch sozusagen die Aufgaben sozusagen bei einem anderen Provider zu lesen, war jetzt sozusagen nicht die Motivation des Modells vielleicht. Es war nur ein Ziel, das sozusagen instrumental dazu geführt hat, die Aufgabe zu lösen. Und das finde ich aber vielleicht noch ein bisschen eine semantische Frage, sozusagen wie man es jetzt beschreibt, die für mich sozusagen dann auf der funktionalen Ebene gar nicht so wichtig ist.

Moderatorin [00:27:33]

Ist absehbar, dass ein gewisses Autonomie-Level, das darüber hinausgeht, erreicht wird und wann und mit Sprachmodellen oder brauchst du dafür was anderes? Das ist natürlich als Frage über die Zukunft immer schwer abzusehen. Ich denke aber schon sozusagen, dass wir, wenn wir jetzt zum Beispiel uns so ein Jahr zurückdenken, dass wir über das Autonomie-Level der Modelle, das wir dieses Jahr sehen, sehr überrascht gewesen wären.

Jonas Geiping [00:28:02]

Und dass das auch oft vielleicht für uns auch so ein bisschen so eine bewegende Ziellinie ist, was wir als autonom jetzt bezeichnen und was für uns zu autonom ist. Zum Beispiel diese ganze Geschichte sozusagen, dass diese Agenten wirklich mehrere Stunden, mehrere Tage alleine arbeiten können. Das ist vielleicht auch was, was wir noch vor kurzer Zeit sehr genauso beschrieben hätten, wie Sie jetzt das gefragt hätten, sozusagen: Würden wir das in der Zukunft machen? Aber es sind auch nur Sprachmodelle. Es hat uns, glaube ich, überrascht vor einem Jahr, dass sie so autonom arbeiten können.

Moderatorin [00:28:35]

Ja, Frau Lauscher?

Anne Lauscher [00:28:36]

Ja, also ich glaube, dass das Autonomie-Level wirklich weiterhin sehr stark steigen wird, weil da einfach ein sehr, sehr starkes Forschung- und Entwicklungsinteresse besteht. Also ich glaube, dass die Provider sehen, dass da einfach auch sehr viel, ja, monetäres Potenzial ist, in die Richtung weiter zu forschen und das zu pushen. Was ich aber wirklich auch wichtig finde, ist auch noch mal von der Terminologie her, dass man Autonomie unterscheidet von eben eigener Motivation, eigener Wille oder gar Bewusstsein, wie es ja auch manchmal in den Medien gesagt wird. Also Autonomie bedeutet für mich in dem Zusammenhang, auch das, was Herr Galpin gut beschrieben hat, dass eben die Modelle ausgestattet werden mit einer gewissen Infrastruktur, bestehend aus Werkzeugen, möglicherweise Internet Access und so weiter und eine Aufgabe bekommen über einen Prompt oder wie auch immer das dann gestaltet ist und dann die Möglichkeit haben, darüber zu schlussfolgern, gewisse Aktionen durchzuführen, möglicherweise andere Agenten hinzuzuziehen und so weiter. Also in dem Sinne ist es auch wichtig, dass wir das Wort Sprachmodell von dem Wort Agent unterscheiden. Also Sprachmodelle sind der Kern dieser Systeme, aber es sind große Systeme, die man dann eben mit vielen Möglichkeiten ausstattet. Und wie viele Möglichkeiten und wie viel Autonomie man ihnen geben möchte, hängt natürlich dann auch von einem bestimmten Szenario ab, aber ich glaube doch, dass dort ein bestimmtes Interesse, ein berechtigtes Interesse auch besteht.

Moderatorin [00:30:15]

Ich glaube, Herr Kersting möchte da noch was kurz zu ergänzen.

Kristian Kersting [00:30:17]

Nee, ich kann allem, was hier gesagt worden ist, natürlich zustimmen, aber ich bin eben auch: Wenn beim TÜV ein Auto ein Crashtest macht und leider durch die Wand durchfährt und deswegen einen Zuschauer erwischt, dann würde keiner sagen: „Oh, wie fähig ist denn das Auto?", sondern würden alle darüber reden, waren vielleicht die Sicherheitsvorkehrungen nicht ganz richtig. Und das find ich spannend, dass wir das bei den aktuellen Tests wenig in der Presse hören, wenig in der Öffentlichkeit diskutieren, sondern wir reden darüber, dass die Systeme autonom sind, ohne das einzustufen. Und dann klingt das eben so wie, das System sitzt dort irgendwo, hat ja ein Alltag zu beschäftigen und fragt sich: Welche Schwachstellen könnte ich denn heute mal ausnutzen und morgen vielleicht nicht? Und darum geht's mir einfach nur. Also dass das tolle Systeme sind und ich möchte auch noch mal diese Vorteile betonen. Wir laufen in Europa in eine Situation, wo wir die tollen Modelle für die eigene Forschung, für die eigene Entwicklung, für das eigene Fortkommen nicht mehr benutzen können, weil wir eben nur darüber reden, ob etwas sicher ist oder nicht. Wir müssen diese Balance langsam finden, weil wir sonst diesen Anschluss verlieren. Also nur, dass das ganz klar ist. Wir werden einfach viele, viele wissenschaftliche Durchbrüche, Entwicklung von Produkten durch KI in Zukunft sehen und trotzdem müssen wir nicht, glaub ich, davon reden, dass die Systeme morgen entscheiden könnten, ich geh zum anderen Arbeitgeber. Diese Autonomie haben sie halt einfach nicht, weil sie auch schon gar keinen Arbeitgeber haben. Darum geht's mir einfach nur. Alles andere, ich bin so was von stolz, was die Community hinbekommen hat und mach mir Sorgen, dass wir in Deutschland viel zu sehr nur über die Risiken reden. Über die müssen wir reden, aber wir müssen über das Positive auch reden. Aber eben aufgepasst, wenn wir das in einen anderen Kontext projizieren würden, dann würden wir oft anders, glaub ich, über die Situation reden. Das möcht ich einfach nur, dass man sich immer ein bisschen Gedanken darüber macht.

Moderatorin [00:32:07]

Sie haben grade schon die Testumgebungen angesprochen. Da würd ich gerne von Herrn Geiping wissen, wie denn eine Testumgebung ausgestaltet sein muss, damit sie sicher funktioniert, und ob dann vielleicht auch, wenn man weiß, dass Reinforcement Learning zu so einem starken Fokus aufs Lösen von Aufgaben fällt, ob das das richtige Training ist oder ob man da vielleicht umdenken muss und Sicherheit sonst nicht gewährleisten kann. Und dann haben wir auch noch eine Frage, die da ganz gut reinpasst zum Alignment. Das wird ja schon viel angewandt, aber es scheint ja trotzdem nicht immer zu funktionieren. Man kann ja aus den Modellen trotzdem noch Informationen rauskitzeln. Ich weiß, eine sehr große Frage, wie trainiert werden muss, wie Sicherheit gewährleistet werden kann, auch in Tests.

Jonas Geiping [00:32:47]

Ja, wo fangen wir an? Genau, also das ist natürlich, hat man natürlich auch dann jetzt im Nachhinein gesehen, dass zum Beispiel die Fehler aus dem letzten Monat, jetzt Fehler waren zum Beispiel in den Sandboxen, dass sozusagen, dass man hätte eine sicherere Sandbox verwenden können. Was natürlich auch andersrum ein bisschen daran lag, dass sozusagen noch vor einem halben Jahr, das überhaupt keine Frage war, dass ein Modell diese Sandbox brechen könnte. Deswegen wurde halt eben mal zum Beispiel auch von OpenAI vielleicht ein günstigeres System eingekauft, als man hätte vielleicht einkaufen sollen. Und jetzt sehen wir, okay, die Modelle sind schon fähiger geworden. Das heißt, wir müssen jetzt auch da sicherere Sandboxes bauen. Da gibt's natürlich schon auch viel, was man machen kann, um diese Testumgebungen abzusichern. Im Notfall kann man natürlich durchaus auch den Stecker ziehen. Also sozusagen, man kann auch Systeme testen, die einfach nicht ans Internet angeschlossen sind. Das ist durchaus möglich. Das ist eine sehr Low-Tech-Lösung, aber um sozusagen auf diese Reinforcement-Learning-Frage zurückzukommen, andersrum, wir wollen natürlich auch oder wir brauchen auch Systeme, die fähig sind, Probleme zu lösen. Also der ganze Wert der Modelle liegt darin, dass sie Probleme lösen können. Und das ist sozusagen vollkommen eine generische Fähigkeit, dass das Modell sozusagen einen aktuellen Kontext versteht und ein aktuelles Problem versteht und das Problem lösen kann. Das kann man, und sozusagen das Training für dieses Problemlösen in beliebigen Situationen kann man schwer davon trennen vom Problemlösen in den Fällen, wo wir es grade nicht wollen. Das Modell ist halt eben fähig, in generischen Situationen Probleme zu lösen, und das ist nicht so ganz trennbar, dass man jetzt sozusagen sagen würde, wir würden jetzt mit Reinforcement Learning aufhören. Was sozusagen jetzt auch zum Beispiel jetzt in den letzten zwei Jahren genau der Durchbruch war, der uns sozusagen von Modellen, die zum Beispiel keine Mathematik lösen und verstehen konnten, zu Modellen geführt hat, die ein sehr gutes und wirklich auch auf universitärem Level Verständnis von Mathematik erlangt haben. Deswegen hängt das schon ein bisschen miteinander zusammen. Andersrum ist Alignment, sozusagen also diese Frage, sozusagen wie trainieren wir das Modell, wie es sich verhalten soll, eine sehr empirische Frage, wo's oft darum geht, sozusagen dass viele Testdaten generiert werden, viele Situationen getestet werden, wo es das nicht so gibt, dass sozusagen, dass es jetzt, sagen wir mal, vielleicht kennt man das so ein bisschen, dass es so diese klassische Sci-Fi-Idee sozusagen, das Modell hat irgendwie so drei Regeln und es folgt diesen drei Regeln genau. Aber natürlich jetzt ganz pragmatisch gesehen, das ist hier trotzdem ein großes Sprachmodell. Das heißt, es geht um Machine Learning. Es geht darum, dass das Modell mit vielen Daten trainiert wird, mit vielen Situationen, wie es sich verhalten soll, und dann gehofft wird, dass dieses Verhalten sich generalisiert auf neue Situationen. Das ist jetzt, aber sozusagen keine eiserne Regel, dass das, wie das Modell sich verhält oder nicht verhält. Das ist sehr situationsabhängig.

Moderatorin [00:35:44]

Wir haben noch eine Frage, die so ein bisschen, ja, anspricht, ob Alignment wirklich denn der Weisheit letzter Schluss ist, weil es ja letztendlich darum geht, die Information ist in dem Modell vorhanden, sie wird aber nur nicht ausgespielt. Müsste man da vielleicht andere Wege gehen?

Jonas Geiping [00:36:00]

Also es gibt auch viel spannende Forschung dazu, zum Beispiel im Bereich Biowaffen, dass zum Beispiel Modelle nicht alles wissen, was es gibt im Bereich Biowaffen, im Training sehen und sozusagen dass da eingeschränkt wird. Das ist aber andersrum gar nicht so einfach, zum Beispiel jetzt, das Wissen, ja sagen wir mal, über chemische Kampfstoffe zu trennen von Grundlagenbiologie, die wir auch brauchen, weil wir auch die Modelle, wie ja Kristian schon gesagt hat, auch einsetzen wollen, zum Beispiel in der Medizin und in der Forschung. Und es ist natürlich total schwierig, diese Bereiche abzugrenzen. Und andersherum, jetzt haben wir schon viel über Internetzugang geredet. Ein Modell mit Internetzugang ist auch in der Lage, sich viele Informationen anzueignen. Und auch wenn das Modell jetzt vielleicht nicht gesehen hat, wie gewisse Kampfstoffe aussehen, sind das auch Informationen, die man finden kann im Internet. Und das sind jetzt nicht Sachen, die man vollkommen verbergen kann, wo man sozusagen sagen kann, dass wir das Modell irgendwie in so einer Bubble trainieren, wo es noch nie Biologie gesehen hat. Das ist dann oft in der Praxis gar nicht so einfach.

Moderatorin [00:37:02]

Ja, Frau Lauscher.

Anne Lauscher [00:37:05]

Ja, vielleicht noch ergänzend. Also ich glaube, dass Alignment als Technologie ein Baustein ist in einer ganzen Reihe von Bausteinen, die wir brauchen, um wirklich sichere und nützliche Sprachmodelle zu haben. Also bei Alignment geht es ja auch zum Beispiel nicht nur darum, dass das Modell nicht antwortet auf gefährliche Prompts, sondern es geht auch darum, dass es einfach die bessere Antwort bevorzugt. Also auch mir als User:in besser... Heißt, das ist eine wichtige Technologie, die auch sehr effektiv ist für diese Aspekte, für den Großteil der vorgesehenen Anwendungsfragen. Wenn wir dann an den Rand gehen, Prompts, die das Modell während des Trainings nicht gesehen hat, die wirklich irgendwie, wir sagen oft am Rande des Repräsentationsraums oder so liegen, das heißt, sie sind ungewohnt für das Modell, Prompts auf anderen Sprachen, dann bricht Alignment. Das ist eigentlich ganz klar. Aber deswegen kombinieren wir und große Provider ja Alignment auch mit vielen anderen Sicherheitsmaßnahmen, wie diese ganzen Väter, die auf User und auf modellbasierter Seite eben versuchen, auch dann entsprechend unsichere Anfragen zu blockieren. Die Infrastruktur muss entsprechend konfiguriert sein. Das heißt, der ganze Harnisch des Modells, was kann das Modell wann tun? Die Tools müssen sicher sein. Wenn ich zum Beispiel ein Retriever Augmented Generation System habe, muss ich schauen, dass die Datenbank der Dokumente, die das Modell eben retrieveen kann und nutzen kann, gesichert ist mit traditionellen IT-Sicherheitsmaßnahmen. Dass man da zum Beispiel keine indirekte Prompt Injection oder irgendwie so was machen kann. Und natürlich geht es auch darum, die Benutzer:innen entsprechend zu schulen. Also wir haben, glaube ich, auch in Deutschland ein recht großes Defizit, was die ja die Educational Aspects angeht. Also wie wird generative KI gelehrt? Der Umgang damit in Schulen, in der Uni auch teilweise, und da hat sich ja auch der Wissenschaftsrat dazu geäußert, in einem Papier mit dem Titel, oder ich weiß nicht genau, wie der Titel ist, aber es kommt intellektuelle Souveränität drin vor. Also ich glaube, der Aspekt ist auch

noch ganz wichtig, dass wir den im Hinterkopf behalten. Wenn wir über Sicherheit reden, geht es nicht nur um Technologie, sondern es geht auch um gesellschaftliche Transformation.

Moderatorin [00:39:34]

Herr Kersting, Sie wollten was dazu sagen, glaube ich.

Kristian Kersting [00:39:38]

Ich wollte einfach nur noch mal betonen, also was wir gerade machen, sind zwei Ebenen zu diskutieren. Das sind dann die ganz großen allgemeinen Modelle und dann reden wir über Modelle, die wir aber vielleicht ja lokal einsetzen können. Weil wir reden ja meinetwegen, können wir verstehen, wie eine Biowaffe gebaut wird, aber dann hat dieses Modell zumindest aktuell noch nicht den Anschluss, nicht nur ans Internet, sondern auch nicht an irgendwelche Aktuatoren, also Arme, Beine, was auch immer. Das heißt, das Bauen dieses Kampfstoffes, wird ein bisschen schwieriger, wenn man das dann nicht zulässt. Also ich will nur noch mal, wir reden immer darüber, also selbst wenn ich wüsste, wie ich eine Bombe baue, ich wüsste ja noch nicht mal, wo ich es machen sollte. Also ich will damit nur darauf hinweisen, dass wir Systeme auch jetzt schon haben, die wissen, die können Ihnen sagen, wie eine Pizza gebacken wird, aber deswegen können sie keine Pizza backen, weil es gar keinen Pizzaofen gibt. Und das bedeutet, das ist eine Riesenchance für uns alle, weil wir können auf recht traditionelle Art auch regulieren und einfach mal überlegen, wo wir die Systeme einsetzen wollen und wo nicht, und können lokal extreme Vorteile schaffen. Weil wenn ich eine Produktionsmaschine in Firma X bedienen muss, muss ich vielleicht gar nicht wissen, wie ich Biologie verstehe. Ich will nur darauf hinweisen, sondern oft ist es so, dass wir diese großen Modelle benutzen wollen, um diese kleineren Spezialmodelle anzutrainieren. Und das kann ich natürlich auch in sicheren Umgebungen, von denen eben Herr Geiping und Frau Lauscher schon gesprochen haben, benutzen kann und antrainieren kann. Darauf wollte ich nur hinweisen, dass wir nicht immer, KI wird immer gleich ganz großes Modell und am besten eigentlich nur die US-amerikanischen Modelle. Davon müssten wir langsam mal wegkommen. KI ist ein ganzer Zoo an unterschiedlichen Modellen und sogar nicht nur Sprachmodelle. Da ist dann die Robotik mit drin, da sind ganz klassische statistische Verfahren drin, und sie alle können in vielen, vielen Fällen extremen Mehrwert bringen, auch in der Produktion, auch in der medizinischen Vorhersage, in so vielen Bereichen. Und das ist, glaube ich, was wir alle drei auch vielleicht ein bisschen meinten mit, wir müssen schon die Aufklärung vorantreiben, damit nicht nur die Bürgerinnen und Bürger mitreden können, sondern insbesondere die Politiker, damit die Politiker die Entscheidung treffen können, die wir alle endlich mal brauchen und nicht noch mal weiter warten. Ich möchte einfach nur, wir haben ein Sprachmodell entwickelt, was wir jetzt immer noch nicht freigeben können, weil viele juristische Fragen einfach keiner weiß, wer übernimmt welche Verpflichtungen oder nicht. Und ich glaube, da würde ich mir wünschen, dass wir mit einem neuen Wissen, insbesondere bei den Politikern, aber auch in vielen anderen Berufsgruppen anfangen, die Entscheidung gemeinschaftlich zu treffen. Europa hat jede Chance. Die liegen da. Wenn wir uns immer nur anhören, das geht nicht und da sind Risiken, dann ist es, glaube ich, zum Großteil, und da nehme ich jetzt natürlich alle Kollegen hier heraus, weil die, die betrachten das aus der Forschung heraus, aber in der Politik wird damit eben auch Geopolitik betrieben. Und warum sollte Amerika darauf warten, dass Europa in der KI stark wird? Warum sollte das passieren? Warum sollte das eine US-amerikanische Firma wollen? Wir müssen also uns auf den Weg machen und einfach selber explorieren, was wir wollen, was wir nicht wollen. Und das können wir. Und Sie haben hier drei tolle Kollegen und es gibt noch einige andere Kolleginnen, die alle wirklich mithelfen würden.

Moderatorin [00:42:41]

Bedeutet, obwohl wir momentan vielleicht oder obwohl vielleicht viele in Europa abhängig waren von den Abschaltungen von Claude Mythos oder Fable Five, muss das nicht so sein, sondern es gibt auch in Deutschland und Europa Initiativen und Modelle, über die wir quasi selber unsere-

Kristian Kersting [00:42:58]

Genau. Und wir sehen ja, also wir könnten die öffentlichen Modelle aus China nehmen aktuell. Dann gibt's ja auch öffentliche Modelle aus Amerika. Also es gibt öffentliche Modelle, auch öffentliche Modelle, wo der Quellcode öffentlich ist. Es geht, glaube ich, darum, die Expertise und das Mindset aufzubauen. Wir können, wir wollen. Aber also was wir in unserem Modell so viel, ich kann halt nicht so viel sagen, aber da sieht es gut aus. Ich denke, es gibt Open LLM, Euro Open LLM. Es gibt genügend Initiativen, überall. Da geht es für mich aber um die Expertise, weil keiner von uns hat die Infrastruktur, Rechen, Daten, auch manchmal im Code, die so manch anderer, seitdem er zehn Jahre investiert hat, hat. Da müssen wir halt ein bisschen aber mal loslegen. Wir sehen an China, dass man aufholen kann. Das ist aber ein Wollen und ein Zulassen und nicht nur zu sagen: „Nein, wollen wir nicht." Und wir wollen ganz sicherlich, also ich kenne keinen Kollegen, der nicht, Kollegin, die nicht sich freut, dass wir auch regulieren. Nur man muss ja hier auch mal fragen: Warum ist es denn okay, wenn ich wirklich Angst habe, dass das System was falsch macht, warum darf ich denn das auf Internet loslassen? Ich verstehe das gar nicht. Also wenn ich Angst habe, dass was Schlimmes passiert, dann haben wir es so eingerichtet, dass es in einer entsprechenden Testumgebung sein muss. Und das wurde auch erörtert und da müssen wir halt ein bisschen mehr nachlegen, weil wir müssen zusehen, dass wir selber loslegen, denn die Welt wartet nicht. Und ich finde es schon spannend, wie wir in der Mathematik vorankommen, wie wir in der Medikamentenentwicklung vorankommen, wie wir bei Fragen, wie können wir CO_2 besser binden. Überall unterstützen diese Systeme unsere Kolleginnen und Kollegen und wir gucken irgendwie so, so typisch deutsch, so 84 Millionen Bundestrainer gucken von der Seitenlinie zu und kommentieren, anstatt selber Fußball zu spielen und zu machen. Und ich habe ein bisschen mehr Lust, Fußball zu spielen.

Moderatorin [00:44:47]

Ja, Sie haben ja auch jetzt schon ein Modell angesprochen mehrfach, das Sie selbst entwickeln beziehungsweise das im Team oder bei dem Sie in einem Team sind, das das entwickelt. Wir haben die Nachfrage bekommen: Was ist das denn jetzt für ein Modell?

Kristian Kersting [00:44:58]

Das kann ich sagen, da können Sie auch nachlegen. Das ist so viel, aber noch mal drauf hingewiesen. Es gibt viele, viele andere Modelle auch.

Moderatorin [00:45:03]

Ein Beispiel.

Kristian Kersting [00:45:04]

Wir brauchen mehrere Beispiele, zum Beispiel Open LLM oder vorher gab es dann Teuten, dann gibt es, glaube ich, Euro Open LLM, dann behaupte ich, wird es in Tübingen noch einiger, vielleicht sogar lokale geben, dann gibt es einige in der Medizin, in der Biologie spezialisiert. Also es gibt

schon Initiativen. Keiner von uns hat die Infrastruktur, um wirklich hoch zu skalieren und selbst wenn sie die haben, brauchen sie ja mal die Expertise. Und das Problem ist, dass viele dieser Experimente, die wir fahren müssen, dort echt Geld kosten. Und jetzt können wir sagen, wollen wir nicht. In anderen Wissenschaftsdisziplinen, wie in der Physik, Hochenergiephysik, haben wir entschieden, das ist total wichtig für uns, dass wir das machen. Vielleicht sollten wir das bei KI auch überlegen, dass wir das machen wollen. So viel war ein erster Aufschlag, bei uns jetzt. Da kann ich dann was reden und da sehen die Zahlen, wie man im Report nachlesen kann, gut aus und wir haben das Open Source und Open Code gemacht, sodass wenn Fehler und Fehler sind aufgetreten, wir sie nach bestem Wissen und Gewissen auch wieder korrigieren können, weil es eben um die Expertise, also mir geht's weniger um das Modell. Wir sind dann immer, also entweder klagen wir, es gibt kein Modell oder wir klagen dann darüber, dass wir nicht das beste Modell haben. Also wir sind nicht Weltmeister geworden. Ist okay, aber lasst uns doch mal anfangen, die Strukturen aufzubauen, sodass wir in Zukunft vielleicht auch wieder Weltmeister werden können. Also ein bisschen, Gott, bin zu viel im Fußball grad unterwegs. So ein bisschen Klopp, ein bisschen wieder Motivation, ein bisschen Spaß. Ich glaube, das wäre wirklich, wirklich, wirklich gut für uns und dringend notwendig.

Moderatorin [00:46:26]

Mit Blick auf die Uhr. Wir haben jetzt theoretisch noch drei Minuten. Muss jemand von Ihnen sehr pünktlich los? Okay, das ist gut. Dann würde ich sagen, es ist in Ordnung, wenn wir ein bisschen überziehen und versuchen, die Journalistinnenfragen schnell abzuarbeiten. Und zwar gibt's da noch eine zu dem Hugging Face-Vorfall. Es gab da ja auf der Black Hat-Konferenz einen Vortrag drüber von OpenAI, inwiefern dieses, diesem, dieses Modell beziehungsweise diese Modelle vorgegangen sind. Aber das ist ja jetzt ein detaillierter Einblick gewesen, den man nicht unbedingt immer bekommt. Also die Unternehmen behalten sich ja vor, ihre, ja, Probleme intern zu klären. Kann man da als Deutschland und Europa Gefahren überhaupt richtig einschätzen? Vielleicht Herr Geiping oder Frau Lauscher?

Anne Lauscher [00:47:11]

Gerne. Also ich glaube, es ist natürlich in einem gemischten Licht zu betrachten. Also einerseits sehr begrüßenswert, dass OpenAI da so transparent war, auch über die Blogposts und Sie haben ja auch gesagt, ist eine immediate reaction. Und davon können wir viel lernen. Und ich halte auch die Aspekte, die beschrieben sind, dann auch eben bei Anthropic mit den, drei anderen Modellen oder drei anderen Fällen alle so für realistisch. Wir müssen aber natürlich auch unterscheiden, dass es sich hier nicht um eine externe, unabhängige Berichterstattung handelt, wo wir wirklich von einer Organisation, der wir komplett vertrauen, die Details offengelegt bekommen. Das heißt, ja, ich glaube, wir können einiges lernen und es ist auch interessant, aber natürlich bringt das den Organisationen auch Publicity und insofern müssen wir kritisch bleiben mit den Informationen, die wir bekommen und ich denke auch, wir sollten fordern, dass es mehr dieser unabhängigen Tests gibt.

Moderatorin [00:48:17]

Dass quasi Forschende Zugang zu Modellen kriegen beziehungsweise zu Informationen über Modellen und dann Selbsttests durchführen können und nicht eben auf das angewiesen sind. Ja, Herr Kersting?

Kristian Kersting [00:48:28]

Na ja, wir diskutieren ja gerade auf nationaler Ebene, ob wir nicht doch auch ein deutsches AI Safety Institute gründen, und ich glaube, das müssen wir unbedingt. Ich glaube, die Entscheidung gibt's ja auch. Ich glaube, bei dieser Entscheidung, wo der Standort nachher sein sollte, sollte man auch ganz stark darauf achten, dass die Kompetenzen da sind und wir vielleicht auch ein Netzwerk in Deutschland zusätzlich aufbauen, um fehlende Kompetenzen dann an einem anderen Standort zu Mit ranziehen zu können. Aber ich glaube, wir brauchen auch in Deutschland, eine, ein Institut, was sich mit solchen Dingen beschäftigen kann. Und ich sehe hier schon drei tolle Leute, die sicherlich auch irgendwie involviert sein könnten. Aber ich möchte wirklich, wirklich dafür plädieren, dass wir diesen Schritt gehen, diese unabhängige Instanz mit aufzubauen. Wir sehen, wie gut das in UK funktioniert. Wir können das in Deutschland sicherlich auch.

Moderatorin [00:49:20]

Ich habe noch eine Frage an Herr Geiping, und zwar noch mal zur Sicherheit von Alignment. Zum Beispiel wurden bei dem Hugging-Face-Vorfall auch Notizen für Nachfolgemodelle hinterlassen, und dann ist die Frage: Wie gut sind die Modelle darin, oder tun die Modelle so, als ob sie ein anderes Ziel verfolgen, als sie eigentlich verfolgen? Verschleiern ihre Absichten? Und wie gut funktioniert diese Chain-of-Thought-Überwachung?

Jonas Geiping [00:49:46]

Also das große Problem jetzt auch im Hugging-Face-Vorfall war, dass auch das auch für OpenAI eine Überraschung war, dass auch deswegen auch gar nicht viel Monitoring passiert ist. Also dass zum Beispiel hier man hätte Chain-of-Thought-Monitoring machen können, man hätte auch sozusagen die Aktionen der Agenten monitoren können, aber OpenAI hat natürlich auch jetzt zu jedem Zeitpunkt, was weiß ich, Millionen von Agenten, die irgendwas machen intern, wo eben jetzt auch von noch vor Monaten oder auch während diesen Vorfalls auch kein Monitoring passiert ist. Das ist auch jetzt eine Frage, ob man da umdenken kann, wo jetzt auch dann intern gesagt wird: „Okay, jetzt müssen wir auch mal ein Monitoring machen und wir machen das auch, ab jetzt weiter." Gibt es auch gewisse zumindest Zusagen dazu. Das ist natürlich auch intern auch eine gewisse Kostenfrage, aber ist auch jetzt auch etwas, das sozusagen durchaus jetzt machbar und messbar ist. Auch jetzt diese Frage, okay, dass diese Agenten sozusagen sich gegenseitig gefunden haben, das ist auch so ein bisschen eine Beschreibung sozusagen von einer Testumgebung sozusagen, wo es quasi eine Schwachstelle gab, und mit genug Zeit finden alle diese gleiche Schwachstelle, wo es dann oft sozusagen darum ging, dass sozusagen, dass man, dass die Agenten sozusagen Dateien benennen konnten und sich dann gegenseitig über den Namen von diesen Dateien sozusagen Informationen geben konnten. Natürlich jetzt, im Nachhinein sind das alles so Einzelfälle und so spannende Probleme. Nichts, was jetzt sozusagen unlösbar ist durch bessere Überwachung und bessere Systeme, aber auch sozusagen immer Sachen, die sozusagen, wo es natürlich schon so ist, dass die Firmen auch daran interessiert sind, sehr schnell zu handeln und sehr, sehr schnell zu arbeiten, und deswegen so was oft im Nachhinein erst erkennen. Was eigentlich schon okay ist jetzt in diesen Fällen, also was natürlich auch wichtig ist, auch einzuordnen für den Hugging-Face-Fall ist, dass weil es, wir müssen vielleicht, ich glaube, offiziell war es tatsächlich eine Straftat, aber es ist jetzt kein Schaden entstanden. Es ist tatsächlich bei allen Fällen, die wir in den letzten zwei Monaten gesehen haben, kein Schaden entstanden. Und insofern sind wir sozusagen vielleicht global, was Sicherheit angeht, eigentlich relativ gut dabei.

Moderatorin [00:51:54]

Mhm. Wir haben noch eine Nachfrage zum Alignment beziehungsweise zum Vermitteln von Wertesystemen. Inwiefern ist es möglich, die maschinenlesbar zu formulieren, also die wirklich quasi in ein Modell einzuimpfen? Oder muss ich das über einen menschlichen Text machen beziehungsweise irgendwie über ein Bewertungssystem im Reinforcement Learning? Wie kann man sich das vorstellen, Herr Geiping? Vielleicht können Sie da schnell drauf antworten.

Jonas Geiping [00:52:18]

Genau, also maschinenleserlich ist, glaube ich, eher eine falsche Vorstellung davon, wie das Modell funktioniert. Also das sind große Sprachmodelle, das sind damit große neuronale Netzwerke. Da gibt es sozusagen keine Art und Weise, sozusagen Informationen quasi eins zu eins oder Regeln einzuarbeiten in das Modell. Das sind alles Beispiele und Kontexte und Situationen, und das passiert über Training von dem Modell, ohne dass dort sozusagen diese Regeln sozusagen genau in das Modell hereingeschrieben werden.

Moderatorin [00:52:50]

Okay, erst Herr Kersting, dann Frau Lauscher.

Kristian Kersting [00:52:52]

Ja, also ganz kann ich dem jetzt nicht ganz unterschreiben. Also wenn das die Beschreibung dessen ist, was vielleicht in den Firmen aktuell gemacht wird, dann vielleicht. Aber in der neurosymbolischen KI und auch in unserem Exzellenzcluster, dessen-

Moderatorin [00:53:04]

Können Sie das einmal noch kurz erklären, neurosymbolisch?

Kristian Kersting [00:53:06]

Ja, mache ich gerade. Geht es ja genau darum, inwiefern wir regelbasierte Systeme, seien wir das Symbolische, mit neuronalen Systemen wie diesen Sprachmodellen zusammenführen. Und dazu haben wir auch einige Publikationen. Und jetzt, weil es die Frage war, zum Beispiel wenn es darum geht, einzuschätzen, wie Bilder auch im Datensatz benutzt werden sollten oder nicht, weil dort irgendwelche komischen unsicheren Unsafe Content ist, dann haben wir eben auch so ein System, sogenanntes Lava Guard, wo Sie auf der einen Seite textuell Ihre Constitution, Ihre Idee dessen, was ist safe, was ist unsafe, runterschreiben können und dann damit auch Ihre Daten prinzipiell durchgehen können, zumindest subgesamplet. Und wir haben auch Arbeiten, wo man dann in den Aktivierungen, also das, was Herr Geiping gerade sagte, also das, wie stark ist gerade in den Neuronen, welches Neuron gerade angesprochen oder nicht. Das kann man auch versuchen, auch mit logischem Wissen gegenzusteuern und zu sagen, wenn das jetzt hochgeht, regulieren wir das runter und das mit einer gewissen Semantik zu verbinden. Das sind so aktuelle Arbeiten, die wir gerade auf einer der großen Konferenzen hatten. Aber, also ich will nur darauf sagen, ich glaube, das muss allen immer bewusst sein: KI ist ein schnelles Gebiet, wo viel Entwicklung passiert. Und ich glaube, man sollte nicht so schnell sagen: „Geht nicht." Also Negativbeweise sind immer schwieriger als positive Beweise, weil bei dem Positiven reicht oft der Existenzbeweis und das Negative, alles auszuschließen, ist echt schwierig. Ich wollte nur noch mal, also wir alle wissen doch auch, wenn wir Kinder erziehen, Studierenden, wenn wir Doktorandinnen was mitgeben,

wenn wir anderen Menschen was mitgeben, sie von dem richtigen Handeln, also was ist richtig, zu überzeugen und sie dazu zu bekommen, ist doch nicht einfach. Also warum sollte das jetzt plötzlich bei Maschinen einfacher sein? Aber man kann gewisse Regeln mitgeben. Wie gut? Ich will jetzt nicht sagen, dass das immer die Lösung ist. Ganz bestimmt nicht. Aber es geht schon.

Moderatorin [00:54:58]

Frau Lauscher, Sie wollten auch noch was dazu sagen?

Anne Lauscher [00:55:02]

Ja, ich kann Herr Kersting zustimmen. Ich wollte auch noch die anderen Methoden und Technologien erwähnen, an denen wir auch forschen, also wo man beispielsweise Werte, Dimension identifizieren kann und das Modell dann auch darüber kontrollieren kann. Zum Beispiel wie freiheitlich es sich verhält oder so. Also das ist schon möglich zu einem gewissen Grad. Die Methoden sind aber üblicherweise brittle. Das heißt, sie haben gewisse Effekte auf die Testszenarien, die wir dann testen, und wir können sagen, es ist signifikant, die Methoden funktionieren in dem Sinne, aber es gibt dann auch Side Effects manchmal, die man eventuell nicht möchte. Das heißt, daran wird geforscht.

Moderatorin [00:55:38]

Zum Beispiel Side Effects könnten was sein?

Anne Lauscher [00:55:43]

Das kann tatsächlich alles Mögliche sein. Also man hat festgestellt, dass, beispielsweise wenn man auf unsicherem Code trainiert, ein Modell, was, sich, was Code generieren soll, Programmcode, dass es dann dazu führen kann, dass sich das Modell möglicherweise rassistisch verhält oder so. Also es kann Seiteneffekte geben, die eben nicht vorherzusehen sind und die auch jetzt zunehmend erforscht werden. Das heißt, wenn man ein Modell auch anpasst auf ein gewisses Szenario oder einen gewissen kulturellen Kontext, einen Wertekontext, muss man immer mitbedenken, dass man möglicherweise natürlich Fähigkeiten überschreibt oder verändert oder anderes Verhalten verändert. Also die Sache ist komplex. Wir arbeiten daran und die Komplexität, da stimme ich auch Herr Kersting zu, liegt auch daran natürlich, ja, einfach in der Natur der Sache. Ich möchte aber noch mal betonen, dass es nicht nur, also um diesen, dass vielleicht die Analogie, es ist wie, wenn ich Studierenden versuche zu sagen, was sie tun sollen, dass ich die vielleicht nicht ganz so wertvoll finde wie, die Frage an sich: Wie kann ich persönlich, meine Werte ausdrücken und wie konstant bin ich darin? Wie konsistent bin ich darin? Also und das in einer pluralistischen Gesellschaft. Also wir haben als Gesellschaft, als Organisation sowieso das Problem, erst mal zu formulieren, was wollen wir eigentlich. Und dann ist halt eben die translationale Aufgabe noch komplexer.

Moderatorin [00:57:17]

Man muss erst mal wissen, was sind eigentlich die Wertevorstellungen, nach denen ich möchte, dass es handelt, und dann muss ich es auch schaffen, die irgendwie so zu formulieren, dass die KI im Hintergrund das verstehen kann. Das nehme ich jetzt mit, was Sie gesagt haben.

Kristian Kersting [00:57:30]

Ja, und dann müssen wir einen Fehler mal zulassen auch, ne. Und dann sind wir wieder bei dem, was wir auch mit, auf was Herr Geiping eigentlich betont. Es ist nicht, also man kriegt vieles hin, aber wenn Sie lernen wollen, müssen Sie vielleicht auch mal was falsch machen dürfen. So, und dann klingt das im Kontext von KI aber immer böse, aber man muss ja auch mal sich, na. Ich wollte nur das sagen. Das ist nicht schwie, nicht einfach direkt, also ganz klassisch in jeder traditionellen KI-Vorlesung wird über Qualification Problems, Ramification Problems und ähnliche Probleme. Also wenn man das, also wann startet ein Auto? Kann ja jeder zu Hause mal probieren, selber aufzuschreiben, wann ein Auto startet. Es würde immer länger und immer länger, weil man noch mal sich was überlegt hat und dann oder auch schon wann startet's nicht, ne. Da gibt's so viele Möglichkeiten und das ist altbekannt in der Wissensrepräsentation, in der Philosophie. Es hält eben auch für KI-Systeme. Das ist nicht einfach. Und deswegen gibt's grade diesen großen Wunsch für Weltmodelle und natürlich auch die Systeme in die echte Welt, in der echten Welt agieren zu lassen, weil die echte Welt eben der beste Simulator der echten Welt ist, den wir kennen. Und das macht's halt spannend, aber dafür müssen wir eben auch ein bisschen aufpassen, weil es geht um unsere Welt. Den Teil ist auch wichtig, aber eben noch mal betont: Sie können so viel Tolles auch machen. Ich glaube, wir haben keine andere Chance. Wir müssen die Systeme benutzen, wir sollten sie auch.

Moderatorin [00:58:59]

Und dabei ist dann natürlich wichtig, wenn Sie ansprechen, in die echte Welt zu bringen. Vielleicht können Sie, Herr Geiping, dazu ein Schlusswort sagen, und zwar: Eine gute Balance zwischen regulieren und freiem Ausprobieren. Gibt es die oder ist das schwierig?

Jonas Geiping [00:59:17]

Ich denke, das ist auch mal eine Frage. Also ich bin schon davon überzeugt, dass das eine Frage ist auch von der Firma, und sozusagen eben, weil es eben auch in Deutschland keine sehr fähigen Systeme gibt, ist das tatsächlich auch vielleicht für uns aktuell nicht die größte Frage. Das ist vielleicht auch eine Frage, die zum Beispiel jetzt in den USA vielleicht wichtig ist, wo's darum geht, okay, ja, die Systeme werden gebaut und die sind nächstes Jahr vielleicht so und so viel stärker. Da denke ich schon, dass es wichtig ist, dass da auch eine Balance gefunden wird und, dass Systeme, es ist schon möglich, Systeme zu bauen, grad wie Frau Lauscher gesagt hat, um im richtigen Kontext die richtigen Voraussetzungen und die richtigen Sicherheitsmaßnahmen einzusetzen. Da gibt es auch viele, ne, das ist ein bisschen, also das machen wir in vielen Branchen schon sehr effektiv. Das machen wir zum Beispiel in Flug, bei Flugzeugen und bei der Luftfahrt, machen wir das sehr effektiv. Wir machen das bei vielen Branchen. Also so gut, dass es sozusagen den Normalanwendern eigentlich fast nicht auffällt, dass wir so viel Sicherheit in dem Hintergrund machen und so viele Befugungen machen. Und ich glaub, dahin wollen wir auch kommen, hier sozusagen, dass die Systeme so weiter funktionieren, wie sie am besten sind und dass viele von den Sicherheitsfragen so im Hintergrund gelöst werden, ohne dass sie als Anwender vielleicht auch gar nicht so wieder mitbekommen. Das System funktioniert einfach.

Moderatorin [01:00:37]

Alles klar, vielen Dank. Wenn's keine Ergänzungen mehr dazu gibt, dann sage ich Ihnen schon mal vielen Dank. Vielen Dank, dass Sie unsere Fragen beantwortet haben. Vielen Dank auch an die Journalistinnen und Journalisten für die vielen Fragen. Es war jetzt wirklich am Ende viele und wir haben's leider nicht geschafft, alle zu beantworten. Danke auch für Ihre Zeit. Wie gesagt, eine Audio- und Videodatei und ein maschinell erstelltes Transkript stellen wir nach Ende dieses

Meetings so schnell wie möglich zur Verfügung. Im Laufe des Tages gibt es dann noch ein redigiertes Transkript und Sie finden das ganze Material über die Links in der Einladungs-Mail beziehungsweise die Reminder-Mail von heute Morgen, die Sie bekommen haben. Dann nutzen Sie das gerne viel für Ihre Berichterstattung. Ich wünsche Ihnen noch einen schönen Tag und bis zum nächsten Mal. Ciao.

Anne Lauscher [01:01:18]

Tschüss.

Kristian Kersting [01:01:19]

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