07.12.2023

Transkript

„AI Act – aktueller Stand und Ausblick“

Expertin und Experte auf dem Podium

- **Prof. Dr. Philipp Hacker**
  Professor für Recht und Ethik der digitalen Gesellschaft, Europa-Universität Viadrina Frankfurt (Oder)
- **Prof. Dr. Sandra Wachter**
  Professorin für Technologie und Regulierung, Oxford Internet Institute, University of Oxford, Vereinigtes Königreich
- **Bastian Zimmermann**
  Redakteur für Digitales und Technologie, Science Media Center Germany,
  und Moderator dieser Veranstaltung

Mitschnitt

- Einen Videomitschnitt finden Sie unter: https://www.sciencemediacenter.de/alle-angebote/press-briefing/details/news/ai-act-aktueller-stand-und-ausblick/
- Falls Sie eine Audiodatei oder eine Sprecheransicht des Videomittschnitts benötigen, können Sie sich an redaktion@sciencemediacenter.de wenden.

Transkript

Moderator [00:00:00]

Hallo, liebe Journalistinnen und Journalisten, herzlich willkommen zu unserem virtuellen Press Briefing zum AI Act. Ich bin Bastian Zimmermann. Ich bin Redakteur beim Science Media Center und ich freue mich, dass ich heute Sandra Wachter und Philipp Hacker hier mit mir dabeihabe. Herzlich willkommen! Schön, dass Sie da sind. Ich stelle Sie gleich noch genauer vor.

Der AI Act ist ja mittlerweile zweieinhalb Jahre in der Mache. Jetzt ist gerade die entscheidende Phase. Gestern, wie Sie alle mitbekommen haben, hat der möglicherweise wichtigste Trilog begonnen. [Er dauert], Stand vor fünf Minuten, immer noch an. Die Pressekonferenz zu den Ergebnissen wurde mehrmals verschoben. Es scheint eine grundsätzliche Einigung zu geben, aber Details sind zum Teil noch unklar. Einige Sachen waren ja auch bis zuletzt nicht klar, zum Beispiel, ob Gesichtserkennung durch KI verboten wird und wie Basismodelle wie GPT-4 reguliert werden sollen. Für die Regulierung der Basismodelle soll es eine Einigung geben laut Medienberichten. Vielleicht können auch Frau Wachter und Herr Hacker gleich etwas dazu sagen. Gerade soll es um die biometrische Gesichtserkennung gehen. Falls wir mitkriegen, dass die Pressekonferenz der EU beginnt, sagen wir auch Bescheid. Aber jetzt sind wir erst einmal hier, um den aktuellen Stand und die wichtigsten Fragen zu besprechen. Eigentlich war ja der Plan, über die Ergebnisse des Trilogs sprechen zu können. Jetzt müssen wir etwas improvisieren. Aber ich denke, das kriegen wir hin.

Wie jedes Mal noch der Hinweis: Über die Frage- und Antwortfunktion von Zoom können Sie Ihre Fragen stellen. Machen Sie davon gerne fleißig Gebrauch. Dann zu unserer Expertin und unserem Experten. Ich stelle Sie einmal kurz vor. Wir haben hier Professor Philipp Hacker. Er ist Professor für Recht und Ethik der digitalen Gesellschaft an der Europa-Universität Viadrina in Frankfurt an der Oder. Und Professorin Sandra Wachter ist Professorin für Technologie und Regulierung am Oxford Internet Institute der University of Oxford im Vereinigten Königreich.

Wir haben einige kurze Eingangsfragen hier für Sie, um den Stand der Diskussion einzuleiten. Herr Hacker, was ist der aktuelle Stand nach dem gestrigen beziehungsweise immer noch andauernden Trilog? Wo steht der AI Act jetzt gerade? Wie geht es dann weiter?

Philipp Hacker [00:02:07]

Vielen Dank. Der AI Act steht jetzt besser da, als wir das gedacht hätten noch vorgestern. Das ist schon einmal die gute Nachricht. Es sieht so aus, [als ob] man hier eine Einigung gefunden [hätte]. Und mich persönlich freut das auch sehr. Ich kann auch nur das wiedergeben, was Sie schon gesagt haben. Foundation Models – da scheint tatsächlich ein Kompromiss gefunden zu sein. Ich werde darauf gleich noch eingehen, vermutlich [basiert der Kompromiss] auf dem spanischen Kompromissvorschlag vom 28. November. Das hieße dann, dass sich die deutsche, französische, italienische Fundamentalopposition gegen die harte Regulierung und die verbindliche Regulierung von Foundation Models nicht durchgesetzt hätte. Ich würde das sehr begrüßen, ehrlich gesagt. Ich war von Anfang an persönlich dagegen. Ich glaube nicht, dass man hier mit Selbstregulierung weiterkommen kann. Wir haben alle gesehen, wenn einer darauf keine Lust mehr hat, Elon Musk zum Beispiel bei Twitter, dann tritt er einfach aus. Und dann gibt es keine Handhabe, diese Modelle noch weiter regulatorisch zu bearbeiten.

Wichtig zu sehen, ist auch, dass wir jetzt den Trilog haben. Da wird es sicherlich heute noch ein Ergebnis geben. Und dann beginnt die eigentliche Feinarbeit. Bis zum 9. Februar, so sagt man, werden dann die Technicals weiter abgehalten. Das sind also die technischen Meetings, wo dann nicht mehr die einzelnen Abgeordneten oder nicht notwendig mit dabei sind, sondern die Sachexperten, die Fachreferenten und -referentinnen, wo dann die einzelnen Schrauben noch einmal angezogen werden beziehungsweise einzelne Dinge feinjustiert werden, auch die Erwägungsgründe noch

einmal genau überarbeitet werden und einzelne Paragraphen oder Artikel noch einmal genauer gefasst werden. Das ist also wichtig zu sehen.

Und ansonsten kann ich gerne kurz was sagen, ganz überblicksartig, zur Regulierung vor allem von Foundation Models [...]. Was man sagen kann, ist, dass es sehr wahrscheinlich ist, dass wir hier zwei große Tiers, also zwei unterschiedliche Systemebenen haben werden der spezifischen Regulierung von Foundation Models, und [...] zwar auf der Ebene der Modellregulierung. Einmal diejenigen Regelungen, die für alle Foundation Models gelten, und dann spezifische Regelungen, die für sogenannte High Impact Foundation Models oder Foundation Models with systemic risk gelten. Die [Regelungen], die für alle gelten – das wird relativ harmlos sein. Da wird man wahrscheinlich sehen: Transparenz – Transparenz über die Trainingsdaten und so weiter, was ja auch sinnvoll und gut ist – und gewisse Regeln über das Urheberrecht. Das war ja schon vorher klar, dass das wohl kommen wird, nämlich dass zum Beispiel ein Compliance-Mechanismus dort eingeführt werden muss. Ein Compliance-Mechanismus, wo das Unternehmen dann aufzeigt, dass entsprechende Vorkehrungen technischer und organisatorischer Art getroffen wurden, um zu verhindern, dass urheberrechtlich geschütztes Material, das man nicht nutzen darf nach europäischem Recht – man darf ja dann ein Veto einlegen als Urheber oder Urheberin –, dass solches Material dennoch verwendet wird. Das wäre jedenfalls das, was auch der spanische Entwurf damals vorgesehen hat. Das [wäre] mir persönlich zu wenig, wenn das so käme. Denn man muss ja sehen, dass auch Foundation Models, die jetzt noch nicht diesen absoluten Topmodellen entsprechen wie GPT-4 oder so, dass auch diese durchaus mit signifikanten Risiken einhergehen können. Und deshalb müsste man da meines Erachtens noch nachsteuern. Aber wenn das jetzt im Trilog nicht kommt, dann wird das schwer. Man bräuchte aber Vorkehrungen für Cybersicherheit, für Inhaltemoderation und auch für AI Safety, Stichwort Strategien, technische Strategien, die verpflichtend sind zur Risikominderung, damit eben nicht Cyber Malware und Viren damit geschrieben werden können, damit Bioterrorismus und auch chemische Waffen damit nicht so leicht, ich will jetzt nicht sagen entwickelt, aber doch die Entwicklung und auch bestimmte tatsächlich terroristische Vorhaben damit unterfüttert werden können.

Das ist also die eine Ebene – und die zweite, dazu ganz kurz, das wäre dann für die High Impact Models. Da zeichnet sich ab, dass man das Ganze am Compute aufhängt, also am Trainingsaufwand. Und hier geistert ja immer die FLOPS-Zahl durch den Raum, also Floating Point Operations [Per Second], die Anzahl der Rechenschritte, die durchlaufen werden, damit ein Modell dann am Ende trainiert ist. Hier geht es um sehr große Zahlen, aber mit denen ist schon auch einiges verbunden. Ursprünglich war vorgesehen von der spanischen Ratspräsidentschaft 10 hoch 26 FLOPS als Threshold, ab dem man dann annehmen kann oder zumindest vermuten kann, dass wir so ein High Impact Model haben. Das könnte sein, dass sich das noch einmal ein bisschen reduziert. Ich persönlich fände es gut, wenn wir auf 10 hoch 24 [hinuntergingen]. Dann wäre nämlich auch klar, dass die jetzigen Modelle, die State of the Art sind, so etwas wie GPT-4, Bard, Gemini, erfasst sind. Wenn wir bei 10 hoch 26 sind, sind die wahrscheinlich nicht erfasst. Das wären dann künftige Modelle, wobei gar nicht klar ist, ob irgendwann einmal überhaupt Modelle diesen Stand erreichen werden. Und ich glaube auch schon [bei den] jetzigen Modellen brauchen wir diese zusätzlichen Kriterien und Anforderungen.

Was sind die? Das wären dann solche Sachen wie Risikomanagement, Red Teaming, zu versuchen Teams auszuwählen, die das System bewusst brechen sollen oder dazu bringen sollen, dass es Dinge tut, die es eigentlich nicht tun soll – und das in einer überwachten Prozedur –, weiterhin dann dort Cybersicherheit und womöglich auch noch Angaben zum Energieverbrauch. Das fände ich auch sehr wichtig, da wird vielleicht Sandra auch noch etwas zu sagen. Und ich kann da später auch gerne noch mehr zu sagen. Die Sustainability Dimension ist ja ganz wichtig bei diesem Thema bei den großen Modellen.

Und dann als Allerletztes noch der große Bereich Open Source. Das ist nach wie vor ziemlich umstritten. Da weiß ich persönlich auch nicht, ob es da jetzt schon eine Einigung gibt. Es ist immer

wieder gesagt worden, dass Open-Source-Modelle grundsätzlich ausgenommen sein sollen vom AI Act, dass sie dann nicht ausgenommen sind, wenn sie in High-Risk-Systeme verbaut werden und dass sie auch für die Foundation-Model-Regulierung nicht ausgenommen werden. Jetzt gab es aber den Vorstoß, die doch eventuell auszunehmen, jedenfalls von den Mindeststandards, nicht von denen für systemische Modelle, aber von denen für Mindeststandards. Da bin ich mal sehr gespannt auf das Ergebnis. Ich würde sagen, eigentlich sollten Open-Source-Modelle, jedenfalls was die Mindeststandards für Foundation Models anbelangt, auch mit drin sein. Denn ob Open Source oder nicht, man kann damit dann eben doch Dinge anstellen, die wir lieber so hier nicht sehen wollen, gerade wenn man an Public Safety Threats und Ähnliches denkt. Das als ersten Input aus meiner Warte.

Moderator [00:09:28]

Vielen Dank erst einmal, Herr Hacker, für die erste Einordnung. Eine Sache, die ich eben nicht gesagt habe: Wir hatten eigentlich geplant, dass wir zu viert heute hier wären. Jan Peters von der TU Darmstadt war eigentlich eingeplant für die fachliche Einschätzung aus KI-Forschungsperspektive. Er ist leider krank. Deswegen sind wir heute nur zu dritt. Deswegen haben wir jetzt keinen, der dezidiert dazu da ist, auch die technischen Fragen nach der Funktionsweise der Modelle einzuordnen. Sie beide befassen sich ja auch ein bisschen damit, wie sie technisch funktionieren und kennen sich damit ja auch ein bisschen aus. Aber Sie stecken nicht selbst in der KI-Forschung drin, die Perspektive fehlt heute ein bisschen. Deswegen fokussieren wir uns eher auf die juristische Einordnung. Aber wir können einfach einmal schauen, wozu Sie sich befugt fühlen, etwas zu sagen und wozu dann eben nicht.

Dann erst einmal noch zu Frau Wachter, zu der Beurteilung des AI Acts. Wie sehen Sie den AI Act bisher, und was sollte Ihrer Meinung nach, Stand jetzt, noch verbessert werden?

Sandra Wachter [00:10:22]

Es ist sehr, sehr spannend. Aber wir werden sehen, wie das Ganze jetzt über die Bühne geht. Es gibt so viele Sachen, die auf meiner Wunschliste draufstehen, wo ich hoffe, dass sich noch etwas tun wird. Das Erste hat ein bisschen mit den Hochrisikogruppen zu tun, was genau da jetzt rein- und rausfällt, da würde ich mir noch ein bisschen Nachbesserung wünschen. Ich finde den Hochrisikobereich grundsätzlich sehr, sehr gut. Das Parlament hat hier vorgeschlagen, weitere Risikogruppen aufzunehmen, also KI, die möglicherweise Wahlen beeinflussen kann, oder eben auch Recommender Systems von großen Plattformen mitaufzunehmen. Das wäre meiner Meinung nach keine schlechte Idee.

Ein großes Hin und Her ist noch mit Emotion AI. Dass das gar kein Hochrisikobereich ist, finde ich sehr, sehr problematisch. [Das] sollte zumindest ein Hochrisikobereich sein. Man kann sich genauso gut überlegen, ob man bestimmte Bereiche nicht "bannen" [verbieten] möchte, also Emotion Recognition, in Education oder im Arbeitsbereich, das sind definitiv Bereiche, wo man sich überlegen kann, dass man das generell verbietet.

Und auch die Frage nach dem Predictive Policing, da ist man ja auch hin- und hergegangen, ob man das einfach "bannen" [verbieten] sollte – meiner Meinung nach eine sehr, sehr gute Idee. Aber das ist alles noch offen. Aber da gibt es noch die Möglichkeit nachzubessern.

Ich bin auch nicht besonders glücklich mit dieser Filterungsidee. Ich habe diese Default-Einschätzung sehr, sehr gut gefunden, dass, sobald man im Hochrisikobereich etwas entwickelt, man automatisch in die Hochrisikogruppe hineinfällt. Jetzt diesen zweiten Schritt zu haben, dann noch einmal sich herauswurschteln zu können, gefällt mir überhaupt nicht. Das bringt große Rechtsunsicherheit mit sich, und ich hoffe, dass man da, selbst wenn man da bleibt, das so klein und "narrow" gestaltet, dass das nicht zu Rechtsumgehung führen kann. Ganz besonders würde ich mich nicht

freuen, wenn man das Ganze mit Self Assessment abhandeln würde. Ich hoffe, dass man das noch wahrnehmen wird.

Ich hoffe auch sehr, sehr stark, dass die ganze Lieferkette verrechtlicht wird. Was bedeutet: Foundation Models und General Purpose AI und auch diejenigen, die es am Ende nutzen, dass alle dort in die Haftung genommen werden. Ich habe das in der Vergangenheit oftmals mit Ton verglichen oder mit Keramik, wo diejenigen, die den Ton zur Verfügung stellen, die Foundation Models, eine Pflicht haben, sicherzustellen, dass der Ton nicht giftig ist, weil sie eben wissen, dass Vasen und Aschenbecher damit geformt werden können, und die, die Vasen und Aschenbecher formen, also Applications darauf bauen, wie ChatGPT, haben genauso Verantwortung wie eben auch diejenigen, die Vasen und Aschenbecher kaufen oder ChatGPT nutzen, um Misinformationen zu betreiben. Alle drei müssen in die Verantwortung gezogen werden. Das ist ganz, ganz wichtig. Das Argument zu sagen, [General] Purpose AI oder Foundation Models können für so viele verschiedene Dinge genutzt werden, dass man die Risiken [gar nicht] vorhersehen kann, ist für mich ein seltsames Argument, weil das würde ja bedeuten, dass es per Definition eine Hochrisikotechnologie ist, wenn ich so viel Dinge damit machen kann, dass ich es unmöglich mehr abschätzen kann. Man muss sich auch die Downstream-Probleme damit anschauen: Wenn das Foundation Model ein Problem hat, dann [strahlt das] auf alle Applikationen [aus]. Und es ist viel, viel wichtiger, dass man gleich bei der Quelle ansetzt. Man stelle sich vor, man hat eine Bergquelle, die Wasser an Haushalte liefert und die ist bleiverseucht und das, was man macht, ist, Filter zwanghaft einzuführen in allen Haushalten, anstatt zur Quelle zu gehen und dort nachzubessern, um sicherzustellen, dass da keine Gifte [eingeleitet] werden.

Moderator [00:14:00]

Das ist interessant, was Sie da gerade sagen, weil genau dazu gibt es auch die erste Frage im Chat. Deswegen hake ich hier direkt ein bei dem Punkt, bevor Sie weitermachen. Die Frage, ob Basismodelle diese vorhandenen Bias eventuell auch vererben und woher Anwender das überhaupt wissen könnten, wenn es in der Lieferkette keine Transparenz gibt.

Sandra Wachter [00:14:18]

Ja, wenn ich [meinen Gedanken noch zu Ende führen darf.] Dann hat man diesen Downstream-Effekt und man sollte wirklich an der Quelle anfangen und auch Wasserfilter einbauen. Man muss beides machen, Application, Foundation Models, weil es eben vererbt werden kann. Das ist sehr, sehr wichtig. Was gemacht werden kann, sind Audits. Ich bin ein großer Fan von Audits. Es ist nur ganz, ganz wichtig, dass wir uns überlegen, welche Art von Audits wir uns da vorstellen. Ganz oft kommt dieses Red Teaming auf, was viele Vorteile hat, aber mit Sicherheit [kein] "Silver Bullet" ist. Das ist auch ein [von der] Industrie [vereinnahmter] Begriff, das ist etwas, das ganz stark von der Industrie propagiert wird. Und das hat auch seine Gründe. Das heißt, die beste Art und Weise, sich Red Teaming vorzustellen, ist, man hat eine Festung oder ein Haus und man möchte wissen, ob etwas gefährlich ist in diesem Haus [...]. Und statt in das Haus hineinzugehen und zu untersuchen, was dort los ist, muss man sich mit dem Türsteher unterhalten. Man kann dem Türsteher bestimmte Fragen stellen. Das sind die API, über die man hineinkommen kann. Und dann kann ich testen. Ich kann fragen: Wie kann ich eine Bombe basteln? Wenn ich eine Antwort bekomme, dann weiß ich, da [gibt es] ein Problem. Oder wenn ich ein Diffusion Model nutze, um Kinderpornografie herzustellen, dann weiß ich, das ist ein Problem. Oder wenn ich hacken, Trade Secrets oder Daten von Menschen herausfinden kann, dann weiß ich, dass es ein Problem gibt. Aber es ist natürlich sehr [eng] zugeschnitten und ich bekomme keinen [Zugang] zu dem ganzen Ding. Das heißt, systematisches Testen würde nicht funktionieren. [Man] müsste eigentlich das testen – unter anderem Red Teaming, das ist schon in Ordnung –, aber eben auch weitere Audits sich überlegen und dann kann man das viel besser abschätzen.

Das Letzte, was ich mir sehr, sehr wünschen würde, das Gegengewicht für die Self-Assessments, die individuellen Rechte. Das Parlament hat ja im Juni vorgeschlagen, dass normale Bürger, Bürgerinnen, vielleicht sogar die NGOs anstelle von Bürgern, sich beschweren können, wenn sie der Meinung sind, dass die Industrie oder die Entwickler sich nicht an diese Regeln halten. Und das würde ein Gegengewicht zu diesen Self-Assessments darstellen, weil die meisten Dinge darauf basieren, dass die Entwickler selbst zertifizieren, ob sie der Meinung sind, dass sie dem AI Act konform sind und den Standards konform sind, die jetzt auch noch über die nächsten ein, zwei Jahre geschrieben werden. Diese Standards sind natürlich hauptsächlich von der Industrie geschrieben. Es gibt nur zwei Zivilgesellschaften, die da mitvertreten sind. Die haben nur Observer-Status, sie dürfen nicht mitstimmen. Und wenn das schon von der Industrie geschrieben ist, man sich selbst zertifizieren kann, ob man diesen Regeln folgt, dann braucht man zumindest einen individuellen Rechtsansatz, der es den Bürgern und Bürgerinnen erlauben würde, auch dagegen vorzugehen. Das ist meine Wunschliste. Wir werden sehen, ob das realistisch ist. Aber das sind so die Hauptdinge, die mir am Herzen liegen.

Moderator [00:17:13]

Vielen Dank erst einmal für die erste Einordnung. Wir haben sogar auch schon einige Fragen hier im Chat. Vielleicht an Sie, Herr Hacker, die Frage, was die sich abzeichnende verbindliche Regulierung für die Entwicklung der Foundation Models bedeutet, ob das ein Hemmnis ist. Was kriegen Sie so aus den Kreisen mit? Halten die Forschenden und vielleicht auch die, die es für Unternehmen entwickeln, [die Regulierung für eine Einschränkung]? Das ist ja gerne das vorgeschobene Argument. Was halten Sie davon?

Philipp Hacker [00:17:38]

Davon halte ich ehrlich gesagt wenig in dem Kontext. Es gibt gerade eine neue schöne Studie, die herausgekommen ist gestern von der Future Society. Und die haben gezeigt, dass, wenn man sich einmal die Entwicklungskosten anschaut für wirklich hochklassige Foundation Models – und wir sprechen hier von zehn hoch 24 FLOPs und mehr –, dann ist man ohnehin bei Entwicklungskosten von ungefähr 60 [Millionen]. *[An dieser Stelle hat Herr Hacker sich versprochen, gemeint war, dass ein solches Modell ca. 60 Millionen Euro kostet, nicht 60 Milliarden; Anm. d. Red.]*. Und die Compliance-Kosten, das haben die auch ausgerechnet, die sich zum Beispiel aus dem Vorschlag des Europäischen Parlaments ergeben würden – und es kann ja sein, dass so etwas Ähnliches jetzt kommt –, die machen dann gerade mal ein Prozent der gesamten Entwicklungskosten aus – ein Prozent. Und das dafür, dass die Systeme dann verbindlich sicherer gemacht werden. Das halte ich für absolut gerechtfertigt und durchaus proportional. Man muss es einfach so sehen: Das sind die absoluten Hochklassemodelle, die wir da sehen. Und wer Champions League spielen will, muss sich halt auch an die Champions-League-Regeln halten und kann nicht sagen: Ich würde dann aber gerne zehnmal wechseln dürfen.

Das heißt, auch kleinere Unternehmen aus Deutschland oder Frankreich, auch Aleph Alpha und Mistral, die können das auch stemmen, solange es mit Praktiken einhergeht, die sowieso Best Industry Practice sind. Und das sind die Regeln, die da vorgeschlagen sind. Es sind keine absolut wilden Dinge, die da verlangt werden, sondern Common Sense, das, was man ohnehin tun würde, genau so etwas wie Red Teaming, Sandra hat es schon angesprochen. Das ist etwas, was die Industrie ohnehin schon macht. Da kann man sicherlich noch weiter daran bohren.

Ich würde zum Beispiel auch sagen, was ganz wichtig ist, wäre ein Access Right für Vetted Researchers. Das haben wir nämlich im DSA [Digital Services Act]: Artikel 40, DSA, sagt: Auf VLOPs, also auf Very Large Online Platforms, dürfen unabhängige Wissenschaftler zugreifen. Das gibt es momentan nicht im AI Act. Ich habe lange auch versucht, dafür zu kämpfen. Ob das am Ende drin ist, bin ich mir nicht sicher. Das ist aber etwas, was absolut fehlen würde und was wir brauchen, noch

eine stärkere, unabhängige Kontrolle. Und wie Sandra sagt, vielleicht nicht nur über APIs, sondern da kann man dann gucken – das ist nicht ganz einfach –, den Researchers soweit [Zugang] zum Modell zu geben, dass sie auch am Modell selber herumbasteln können. Man muss natürlich aufpassen, dass dabei nicht gleich Trade Secrets dann rausgeleakt werden, sonst kommt halt irgendein Researcher, der letztlich aus China kommt und sagt: Ich hätte auch gerne Zugang, und flups, dann ist auf einmal das Modell kopiert. Das muss man definitiv verhindern. Aber da kann man noch einiges machen.

Moderator [00:20:12]

Frau Wachter, Sie hatten eine Ergänzung.

Sandra Wachter [00:20:14]

Philipp Hacker hat mir schon aus dem Kopf herausgesprochen: Man kann ganz, ganz stark Inspiration vom Digital Services Act nehmen, wo wir diese Vetted Researchers haben. Das fände ich ganz gut, weil, wie gesagt, Red Teaming ist eine Methode, aber in Wirklichkeit braucht man viel stärkere Möglichkeiten, mit den Audits hineinzugehen. Der DSA hat ein paar Grundlagen, wie das geregelt werden soll. Da kommen noch Standards raus. Daran könnte man sich sehr schnell orientieren, das [hielte] ich für eine sehr gute Idee. Und nur zu der generellen Frage: wirtschaftlicher Nachteil. Man muss sich nur einmal überlegen, das würde in der Autoindustrie vorkommen, und jemand sagt, ich würde gerne Autos auf dem europäischen Markt zulassen. Aber das mache ich nur, wenn ich mich an keine Geschwindigkeitsbeschränkungen halten muss, wenn ich keine Airbags drin haben muss und wenn ich keine Gurte einbauen muss und wenn meine Autofahrer auch keinen Führerschein haben müssen. Dann würde ich sagen, das ist verrückt, das kann man doch nicht machen. Und das ist genau so, wie es beim AI Act auch ist. Wenn man ein Modell hat oder eine Applikation, die grundsätzlich risikoträchtig ist, dann sind die Dinge, die dann hineingeschrieben sind, nicht [zum Vergnügen] oder um jemanden zurückzuhalten oder Forschung aufzuhalten oder Profit aufzuhalten, sondern um sicherzugehen, dass Leute [technischer Aussetzer] oder verletzt werden. Das ist einfach eine ganz vernünftige Regelung der guten Praxis, um sichere und menschenrechtsschützende Technologien auf den Markt zu bringen. Wie gesagt, das Recht ist dazu da, um als eine Leitlinie für gute Innovation und sichere Innovation da zu stehen.

Moderator [00:21:55]

Wir haben hier eine Nachfrage zu den Open-Source-Modellen. Herr Hacker, vielleicht an Sie, weil Sie sie im Eingangsstatement angesprochen hatten, dass sie den Verhandlungsunterlagen zufolge von der strikten Regulierung ausgenommen werden und die Regulierung dann nur bei der Einstufung als Hochrisikoanwendung oder einer Nutzung für verbotene Zwecke greifen würde. Und da ist die Frage, was Sie davon halten, wenn das denn stimmt?

Philipp Hacker [00:22:17]

Ich halte es für sinnvoll, einerseits zu sagen, dass Open-Source-Modelle grundsätzlich Teil des europäischen und auch sonstigen Ökosystems sein sollten, weil klar ist, dass wir sonst noch stärkere Tendenzen hin zu einer Oligopolisierung des Marktes haben. Im Grundsatz, glaube ich, sind viele Leute auch gerade aus dem Tech-Bereich, die Tech-affin sind, große Open-Source-Fans. Und in ganz vielen Bereichen werden Open-Source-Modelle ja auch eingesetzt, auch jetzt schon in allen möglichen Formen von kommerzieller Software. Zugleich sind diese Modelle so gestrickt, dass da natürlich tatsächlich manchmal einfach Developer Teams dahinterstehen, die das eher hobbymäßig machen. Die können keine großen Compliance-Anforderungen schultern. Aber das sieht anders

aus, wenn sie in kritischen Bereichen eingesetzt werden. Das heißt, da wäre ich dann auch sehr dafür, dass tatsächlich Open-Source-Modelle auch miterfasst sind, nämlich dann, wenn es in die Hochrisikobereiche oder in die verbotenen Bereiche gar geht. Aber verboten ist dann verboten. Ob Open Source oder nicht, das kann natürlich keine Rolle spielen. Und wichtig wäre mir, dass eigentlich alle Foundation-Model-Regeln möglichst auch für Open-Source-Modelle gelten. Denn auch da gilt das, was Sandra Wachter gerade gesagt hat: Ich kann nicht sagen: "Ich bin ein Start-up und mache das Ganze aus altruistischen Motiven." Deshalb kann ich jetzt ein Auto ohne Airbag auf den Markt bringen. Das macht einfach keinen Sinn. Also bei diesen Modellen, die so wirkmächtig sind, da muss man das dann auch zusätzlich fordern können.

Und let's face it: Die Entwicklungskosten sind ohnehin so hoch bei diesen Foundation Models, das macht nicht irgendeine Bude in der Garage oder irgendwo im Keller. Das sind hochprofessionelle Einheiten. Wenn wir sehen, wer das macht, das macht Meta. Also Meta AI ist da ganz vorne mit dabei. Und Falcon, das sind die Leute aus den Vereinigten Arabischen Emiraten. Beides nicht gerade Unternehmungen, die dafür bekannt wären, dass sie an Geldmangel leiden. Insofern, da habe ich ehrlich gesagt relativ wenig Verständnis dafür, wenn man hier jetzt Open Source Models ausnehmen würde.

Ich würde sogar persönlich noch weitergehen, auch wenn das vielleicht keine populäre Forderung ist. Aber ich würde sogar sagen, wenn wir im Bereich Systemic Risk Models sind oder im Bereich 10 hoch 24 FLOPS und mehr, dann würde ich sogar verbieten, diese Modelle Open Source zu stellen. Warum? Weil Forschung gezeigt hat, dass Security Layers, die eingezogen werden, relativ schnell und sehr einfach mit wenig Aufwand entfernt werden können, sobald das Modell Open Source gestellt ist, ich es auf den eigenen Computer herunterladen kann. Das heißt, dann nutzen unsere ganzen Vorgaben, die wir vorher machen, mit Red Teaming und was weiß ich, das nutzt alles nichts mehr, wenn das Modell Open Source gestellt wird, weil das dann einfach "reverse engineered" werden kann. Dann kann ich das eben nutzen, ohne diese internen Sicherheitsbeschränkungen. [...] Wir reden hier letzten Endes über Dual-Use-Güter. Das ist nicht ungefährlich. Und deshalb brauchen wir hierfür eine Lösung, wo wir sagen: regulated oder moderated access in jeglicher Hinsicht. Access für die Researcher, aber gleichzeitig eben mindestens oder höchstens Access, also Hosted Model, eben kein Open Source in diesen ganz hohen Performanzbereichen.

Moderator [00:25:30]

Eine Frage an Sie, Frau Wachter, die im Chat auch kam, ist, warum sich Deutschland, Italien und Frankreich gegen die verbindliche Regulierung der Foundation Models positioniert haben und, noch von mir hinzugefügt, was halten Sie von dieser Position?

Sandra Wachter [00:25:48]

Grundsätzlich ist es nicht überraschend, dass zehn vor zwölf noch einmal Versuche gemacht werden, das eine oder das andere herumzurücken und dann noch einmal das Ganze zu beeinflussen. Grundsätzlich bin ich der Meinung, dass da wahrscheinlich sehr starke Lobbying-Strategien gefahren worden sind, dass dann kurz vor dem Ende noch einmal gesagt wurde, nein, wir sollen jetzt die Foundation Models doch nicht regulieren. Das hat mich persönlich sehr, sehr enttäuscht. Ich hätte mir gedacht, dass wir den Weg bereits gegangen sind und dass das Ding nicht noch einmal aufgemacht wird. Das halte ich für enttäuschend. Man muss sich auch überlegen, welche Risiken es gibt. Hunderte Millionen Leute, die in Europa leben, deren Grund- und Menschenrechte potenziell eingeschränkt, verletzt werden. Von Diskriminierung zur Misinformation, zum Umwelt-Impact, den diese Dinge haben. Und der Vorschlag, das mit Selbstregulierung zu machen, ist eigentlich fast empörend. Weil Self Regulation bedeutet nichts anderes als: Ich kann mich an die Regeln halten oder auch nicht. Im juristischen Fachjargon nennt man das Lex Imperfecta, unperfektes Recht, wenn es keine Möglichkeit gibt, es durchzusetzen, wenn sich jemand daran nicht halten möchte. Und

deswegen verstehe ich die Logik dahinter überhaupt nicht, denn den einzigen [Anreiz], den ich schaffe, ist, mich nicht daran zu halten. Deswegen würde ich es für wesentlich besser halten, ordentliche Rechtsvorschriften zu haben.

Philipp Hacker hat auch schon gemeint, wie wir das mit Elon Musk und Twitter gesehen haben, dieser Code of Conduct, den es für Misinformationen gegeben [hat], da hat er sich auch einfach nur [darüber hinweggesetzt]. Das kann man so nicht machen. Wir sind im Hochrisikobereich, wir [haben] Foundation Models, und wir wissen, dass diese Risiken da sind. Und sich auf das ethische Gefühl eines Unternehmens zu verlassen, ist einfach nicht genug. Man muss Nicht-Compliance unter Strafe stellen.

Moderator [00:27:50]

Herr Hacker, Sie hatten auch noch eine Anmerkung dazu.

Philipp Hacker [00:27:53]

Vielleicht auch nur ganz kurz zum Hintergrund der deutsch-französisch-italienischen Position. Ich glaube, dass da zwei Sachen auch eine Rolle gespielt haben. Das eine ist, dass natürlich Deutschland und Frankreich tatsächlich die einzigen Staaten sind, die selbst hoffnungsvolle Foundation Model Provider haben. Insofern ist es sicherlich kein Zufall, dass gerade diese Staaten sehr dafür sind, die Anforderungen niedrig zu halten. Das ist aber meiner Ansicht nach schon ökonomisch eine Milchmädchenrechnung, denn wir haben hier sehr wenige, durchaus starke Unternehmen, Aleph Alpha, Nyonic in Deutschland, Mistral, Silo in Frankreich.

Aber das sind eben auch nur zwei Unternehmen hier und zwei da und vielleicht noch ein paar weitere. Aber der ganz große Teil des europäischen Ökosystems im KI-Bereich sind eben keine Foundation Model Provider, sondern sind andere Unternehmen, die entweder auf Foundation Models aufbauen oder ganz andere KI-Modelle nutzen, zum Beispiel Convoluted Neural Networks, um damit Krebserkennung zu machen. Und das Problem ist jetzt, wenn ich die Foundation Model Provider ausnehme und sage, och, die müssen nur ein bisschen Selbstregulierung machen und wer das nicht will, der macht das dann am Ende des Tages halt nicht, dann wird ja die Regulierungslast notwendig auf die nachgelagerten verschoben. Und das ist eben das ganz große Ökosystem, die sich dann nicht darauf verlassen können, dass sie da eigentlich gute Qualität geliefert bekommen. So wie Sandra Wachter das schon sagte, macht es überhaupt keinen Sinn, nicht einmal an der Wurzel, sondern tausendmal in der Anwendung den gleichen Fehler auszumerzen.

Das heißt, das ist schon mal das eine, dass man sich ökonomisch vergriffen hat, weil man nicht gesehen hat, dass das eigentliche Zentrum des KI-Ökosystems in Europa eben gerade nicht Foundation Models sind bislang. Das ist leider so, aber daran wird sich so schnell auch nichts ändern. Deshalb ist das aus dem Grund schon einmal wirklich seltsam.

Was man eigentlich damit machen wollte, vermute ich, ist das Narrativ zu ändern. Zu sagen: Schaut mal, vor allem liebe VCs, also liebe Venture-Capital-Leute, wir in Europa, wir wollen KI. Wir wollen hier noch einmal ein Zeichen setzen, dass wir nicht nur für Regulierung oder Überregulierung stehen. Und ich meine, so ein Narrativ ist ja an sich auch richtig. Ich muss auch sagen, es besteht ja auch ein Risiko darin, KI nicht zu nutzen, gerade in medizinischen Kontexten. Man muss es alles gut bauen, und Sandra Wachter hat tolle Paper dazu geschrieben, was alles schiefgehen kann, aber trotzdem haben wir da eine massive Unterversorgung, und es gibt eben schon einige Felder, wo KI mindestens genauso gut oder teilweise besser ist als die Menschen in der Erkennung spezifischer Krankheiten. Und das sollten wir und müssen wir auch nutzen. Auch im Education-Bereich, wo wir auch ein massives Fachkräfteproblem haben. Aber das Ganze geht wirklich nur, wenn wir das insgesamt verantwortungsvoll und sozial und auch ökologisch nachhaltig gestalten.

Trotzdem: Dieser Versuch, das Narrativ zu ändern, der ist vom Impetus her schon richtig. Nur wir sollten es anders machen, indem wir sagen: Guck mal, es gibt hier bestimmte Förderungen, es gibt Safe Harbours. Wir spezifizieren also bestimmte Bereiche auch technisch, wo man sagen kann, wenn sich die Unternehmen daran halten, dann seid ihr im sicheren Bereich, dann seid ihr in der grünen Zone und dann kann euch erst einmal rechtlich nicht wirklich etwas passieren.

Moderator [00:31:11]

Da gibt es direkt eine Nachfrage, die dazu passt. Und zwar, wie wichtig eigene Foundation Models in Deutschland beziehungsweise in Europa sind. Darüber haben wir beim Press Briefing letzte Woche auch schon gesprochen, und Holger Hoos und Kristian Kersting sagten, dass es schon sehr schön wäre, wenn man nicht von den internationalen Anbietern so abhängig wäre. Sehen Sie das ähnlich, oder was ist Ihre Einschätzung dazu? Frau Wachter, fangen Sie gerne an und dann Herr Hacker.

Sandra Wachter [00:31:39]

Natürlich wäre das gut. Es ist total wichtig zu erkennen, dass, wie wir schon gesagt haben, diejenigen, die sie entwickeln, entweder in China oder Amerika [sitzen]. Das sind die großen Player. Man muss sich überlegen, dass je mehr Applikationen auf diesen Foundation Models gebaut werden, sich eine Abhängigkeit ergibt. Das darf man natürlich nicht [vergessen]. Wie gesagt, [...] wenn die Wasserquelle weg ist, dann hat niemand mehr etwas zu trinken zu Hause. Da ist ein gewisses Abhängigkeitsverhältnis, das geschaffen wird. Und wenn das etwas ist, das nicht im Haus passiert, dann ist es natürlich [problematischer, weil] man dann an andere Gewalten gebunden ist. Natürlich. Deswegen wäre es fantastisch, wenn wir Ähnliches hier schaffen könnten. Und die Antwort ist wie immer: Geld. Es geht darum, dass man genug in die Forschung investiert, nicht nur in die Industrie, [sondern] auch in die Wissenschaften, Universitäten, denen Ressourcen zur Verfügung stellt, um dort mithalten zu können, um eben Alternativen bilden zu können. Das ist total wichtig. Aber wie gesagt, es ist immer eine schwierige Sache, das nach außen zu verkaufen, und dass man sagt, man braucht ein bisschen mehr Geld, um mithalten zu können oder zumindest Alternativen zu finden. Ob man in Europa mithalten kann mit Meta, ist eine andere Frage. Aber man kann zumindest etwas schaffen, das eine Alternative [bildet], sodass man nicht nur von dieser einen Wasserquelle abhängig ist. Aber es ist eine Frage vom lieben Geld – wie so oft.

Moderator [00:33:10]

Ja, Herr Hacker noch.

Philipp Hacker [00:33:12]

Dazu vielleicht ganz kurz: Es werden ja tatsächlich auch Anstrengungen unternommen seitens einiger Unternehmen. Ich wünsche denen auch wirklich alles Gute. Das wäre großartig, wenn wir das sehen, dass hier stärker auch Foundation Models auch in Europa ansässig sind. Sie kennen ja vielleicht die Zahlen aus dem Jahr 2022. 73 Prozent der Modelle [wurden] in den USA veröffentlicht, 15 Prozent in China, wahrscheinlich noch massiv "underreported", weil da auch sehr viel im Militärbereich passiert. Das heißt, wir haben weniger als zehn Prozent in Europa, weil auch in den Vereinigten Arabischen Emiraten und dem Rest Südostasiens ein bisschen was passiert. Das ist bedrohlich.

Warum ist das bedrohlich? Ich möchte noch eine Komponente hinzunehmen, nämlich die geostrategische. Ich habe gerade einen Talk gehalten bei einer EuroDefense-Tagung, wo auch viele

hochrangige Generäle da waren. Und da sieht man das als ganz erhebliche Gefahr. Denn wir sehen ja, dass auch militärisch KI immer stärker genutzt wird, sowohl seitens der NATO, als auch von den anderen Staaten, sagen wir mal einem strategischen Rivalen. Und wenn man sich jetzt diese Zahlen anschaut, USA und China als die einzigen Partner bisher, dann muss man sagen, China ist nur sehr bedingt verlässlich. Aber auch die USA sind leider kein wirklich verlässlicher Partner mehr. Wir laufen hier sehenden Auges in die nächste Öl- und Gasabhängigkeit sozusagen hinein, nur diesmal noch schlimmer eigentlich, nämlich in der Schlüsseltechnologie des 21. Jahrhunderts.

Und warum ist das problematisch mit den USA? Wenn man sich die Umfragen anschaut, müssen wir ernsthaft davon ausgehen, dass ab dem Jahr 2025 eventuell Trump wieder an der Macht ist. Und wir haben gesehen, dass der einfach alles nutzt, völlig rücksichtslos, um seine politische Agenda durchzuziehen. Es könnte also durchaus sein, dass Trump dann sagt: Ja, also entweder ihr stellt jetzt die Ukraine-Hilfe ein oder wir stellen den Zugang zu den amerikanischen Foundation Models ein, dann könnt ihr mal gucken, wo ihr bleibt. Und dafür müssen wir gewappnet sein. Das heißt, wir brauchen dringend Player hier.

Aber wie Sandra Wachter schon sagte, da ist es nicht damit getan, dass man ein paar Millionen hier oder da investiert. Deutschland hat es jetzt noch nicht einmal geschafft, die LEAM-Initiative zu unterstützen. Da geht es um 200 bis 300 Millionen. Das sind, ehrlich gesagt, lächerliche Beträge in dieser Liga. Wir haben gerade zehn Milliarden investiert, oder zumindest versprochen für eine Chipfabrik. Angesichts der Haushaltslage ist das jetzt ein bisschen schwierig in Deutschland. Aber was wir eigentlich bräuchten, ist ein europaweites Paket, ein Investitionspaket, so etwas wie ein europäisches DARPA, ein europäisches Institut, das sich wirklich den Schnittbereichen von AI Research, teilweise auch militärischer Forschung und diesen Foundation Models verschreibt, wo ein paar Milliarden reingepackt werden. Das würde wirklich einen Unterschied machen, denn Aleph Alpha hat gerade 500 Millionen eingesammelt. Das ist hervorragend. Aber wenn man es vergleicht mit den 13 Milliarden, die allein Microsoft schon in OpenAI investiert hat, dann sieht man auch, dass man da sonst sehr schnell an die Grenzen kommt.

Moderator [00:36:03]

Das mit der Abhängigkeit ist auch ein wichtiger Punkt. Und ich glaube, gerade jetzt bei den kürzlichen OpenAI-Dramen haben sich auch einige Nutzerinnen und Nutzer gedacht, ob es jetzt so sinnvoll ist, von dieser Firma abzuhängen, wo es manchmal ein bisschen merkwürdig zugeht. Apropos merkwürdig, Frau Wachter, eine Frage an Sie zum generellen Prozess. Einige der EU-Abgeordneten haben ein bisschen stolz getweetet, hey, wir verhandeln seit 16 Stunden. Kommt bei solchen Verhandlungsmarathons, bei denen am Ende sich die Person durchsetzt, die am meisten Kaffee getrunken hat, am Ende überhaupt eine sinnvolle Regulierung heraus oder ist der Prozess grundsätzlich schon zu kritisieren?

Sandra Wachter [00:36:48]

[...] Ja, natürlich ist es besser, wenn jeder zu einer vernünftigen Zeit nach Hause gehen und schlafen und dann frisch und munter wieder anfangen [kann]. Aber tatsächlich stehen die Leute ja auch ein bisschen unter Zeitdruck. Wie gesagt, ob [man] sich jetzt vor Weihnachten noch einmal einigt, wäre ja ganz gut. Wenn das jetzt mit den Neuwahlen wieder in Verzug gerät, besteht die Gefahr, dass das Ganze überhaupt nicht mehr durchkommt. Da ist wahrscheinlich einfach ein großer Druck, das einfach durchzudrücken auf Biegen und Brechen, weil man eben nicht die gesamten letzten zwei, drei Jahre verlieren möchte, und die Drafts, die davor gewesen sind, dass Verhandlungen angefangen haben, und das dann vielleicht für ewig wieder liegenzulassen, das verstehe ich schon. Natürlich, wenn man jetzt sich nicht einigen kann und dann wartet man eben bis Anfang Jänner, dann ist auch nichts vertan. [Es ist] viel besser, eine ordentliche Regelung zu haben, [mit der] Grund- und Menschenrechte geschützt werden. Und ein bisschen zu vertagen und das einfach im

Jänner zu machen, sollte auch auch kein Drama sein. Hoffen wir mal, dass es schmerzfrei jetzt noch über die Bühne geht. Aber notfalls ist es auch kein Problem, wenn das Ganze erst im Jänner käme.

Moderator [00:38:04]

Herr Hacker, ich hab gesehen, Sie sind gerade schon dabei, auf die eine Frage einzugehen. Die hatte ich nämlich jetzt als nächste für Sie präpariert. Das ist eine Kritik an einem Kriterium der fürs Training verwendeten Rechenleistung, dass zukünftige Architekturen vielleicht einfach effizienter sind und dann die Foundation Models mit weniger Rechenleistung auskommen und dennoch eine bessere, gute Leistung liefern. Und dann die Frage: Dieses Kriterium, die dann nach der Rechenleistung oder der verwendeten Rechenleistung zu regulieren, ist das überhaupt sinnvoll und zukunftsfähig? Jetzt ist Ihr Ton weg, oder nur bei mir?

Sandra Wachter [00:38:46]

Ich konnte ihn auch nicht hören.

Moderator [00:38:50]

Sie sind allerdings nicht stumm. Vielleicht müssen wir noch mal bei dem Mikro gucken. Ich würde sagen, Sie gucken noch einmal schnell, und wir gucken vielleicht mit der nächsten Frage an Sie direkt weiter, Frau Wachter. Schließt vielleicht auch an an die, die ich eben gestellt habe. Haben Sie den Eindruck bei der Beratung zu diesem Gesetz wurden Wissenschaft und Expertise genug gehört? Natürlich ist es irgendwann so, dass es in die Lobbyschlachten übergeht. Aber haben Sie den Eindruck, das ist zu wenig gewesen, oder war das gut?

Sandra Wachter [00:39:21]

Ich hatte das Gefühl, es ist leider wirklich zu wenig gewesen. Wissenschaft ist an und für sich ein guter Ansprechpartner, weil die die Möglichkeit haben, neutral zu sein, ihre Forschung unabhängig vorstellen zu können.

Und gerade bei dem Beispiel, das Sie genannt haben mit dem Compute Power, ist das jetzt wirklich sinnvoll, das an der Compute Power anzuhängen? Es gab ja auch den Vorschlag, das nicht so zu machen und auf die systemischen Risiken zu schauen, was ich [...] eine sehr, sehr gelungene Idee fände, wenn man sagt, es ist mir egal, wie groß und mächtig das Ding ist und wie viele Daten es verschluckt, wenn es dazu beitragen kann, dass es zu Misinformation beiträgt oder dazu beiträgt, dass massenhaft "reputationally damaging information" verbreitet wird oder Diskriminierung massenhaft stattfindet, all diese Dinge, dann [muss man sich mit diesem Problem auseinandersetzen]. Dann ist mir egal, ob das auf vier oder fünf oder Trilliarden von Daten trainiert wurde. Wenn das Risiko da ist, dann muss man sich eben da etwas überlegen. Es ist eine hilfreiche Art und Weise darüber nachzudenken, weil es oftmals so ist, dass wenn viele Ressourcen benutzt werden, dass der Umwelt-Impact dann größer ist, ganz klar. Und dass wenn mehr Compute drin ist, das es auch zu mehr Bias führen kann. Aber das heißt nicht, dass kleinere Modelle weniger gefährlich sind. Somit kann man da wahrscheinlich einen Mittelweg finden und sich nicht nur an dem Compute aufhängen, sondern an den systemischen Risiken, die [damit] einhergehen, ganz egal, wie groß das Modell in Wirklichkeit ist.

Moderator [00:40:58]

Da sind Sie auch schon auf die Frage eingegangen. Herr Hacker, wie sieht es bei Ihnen aus?

Philipp Hacker [00:41:02]

Hören Sie mich?

Moderator [00:41:04]

Ja, ja. Sehr gut. Wenn Sie noch etwas zu der Frage hinzufügen wollen, gerne.

Philipp Hacker [00:41:08]

Ja, gerne. Das grundsätzliche Problem ist, dass nicht nur das Recht, sondern auch die AI-Forschung selber von den Ereignissen überholt wurde, von ihrem eigenen Erfolg. Und das, was hinterherhinkt, ist die Möglichkeit, diese Modelle vernünftig zu testen und zu evaluieren. Das sagen einem auch die Forscherinnen und Forscher, die genau an den Modellen arbeiten und die mit ihnen Riesenprobleme haben. Wir hatten neulich ein Gespräch mit einer hohen Abgesandten von OpenAI, die genau an solchen Schnittstellen forscht und die selber sagt, sie verstehen es halt nicht vollständig. Das ist ja auch vollkommen klar. Die Modelle sind extrem komplex, und das ist ein Riesenproblem. Da muss man einerseits sagen, wir brauchen mehr öffentliche Investitionen in diese Sicherheitsforschung, um diese Test-Szenarios einfach besser zu machen.

Aber solange es das nicht gibt, solange man nicht gut sagen kann, welches Modell ist jetzt hochperformant und kann deshalb besonders gut bestimmte gefährliche Aspekte verarbeiten – zum Beispiel neue Formen von DNA-Sequenzen oder RNA-Sequenzen für neue Viren oder bestimmte Chemikalien generieren, das geht mit besonders potenten Modellen besser als mit weniger potenten –, solange müssen wir sogenannte Proxys nehmen. Also müssen wir gucken, dass wir irgendwelche Metriken finden, die dieses Gefährdungspotenzial annähernd beschreiben können. Und da ist man in der internationalen Forschung schon der Meinung, dass diese Compute-Zahlen zumindest ein Kriterium sein können.

Aber es ist ganz richtig, es ist womöglich in einem Jahr schon wieder ganz anders. Deshalb ist es aber auch so und das macht Sinn, dass bisher nach allem, was ich weiß und was ja auch von der spanischen Ratspräsidentschaft in den Überlegungen drin war, dass diese Compute-Zahl nur ein Anhaltspunkt ist. Das heißt, wenn wir eine bestimmte Compute-Zahl überschritten haben, dann wird vermutet, dass es sich um so ein Hochrisiko-Foundation-Model handelt. Es können aber auch Modelle aufgenommen werden, die weniger Compute haben. Da bin ich ganz bei Sandra Wachter.

Es kommt am Ende des Tages drauf an, was diese Modelle für Schaden anrichten können, nicht mit wie viel Compute sie trainiert wurden. Und umgekehrt ist es auch so: Der Foundation Model Provider kann im Einzelfall nachweisen, dass er zwar sehr viel Compute gebraucht hat, aber dass keine systemischen Risiken, warum auch immer, [von ihm] ausgehen. Aber eine Korrelation – das hat Sandra Wachter ja auch schon gesagt – ist natürlich klar: Je mehr Compute desto mehr Emissions. Das heißt, letztlich sollten wir darauf hinwirken, mit einer Reihe von Policy Measures, die zum Teil eben noch nicht im AI Act stehen, zum Beispiel Sustainability Impact Assessments. Letztlich kann [auch nachdenken über] die Einbeziehung von AI-spezifischen Komponenten in den europäischen Emissionshandel [und] dafür sorgen, dass diese Modelle insgesamt weniger Energie verbrauchen.

Moderator [00:44:08]

Da kam auch eine Nachfrage direkt zu den Small Language Models, die die Zukunft sein könnten statt Generalistenmodellen. Das haben wir ja auch bei Gemini jetzt gesehen, dass sie diese Nanoversion präsentiert haben, die auf den neuen Google-Pixel-Handys laufen soll, die dann vermutlich vom Compute her weniger hatte. Sie haben schon gesagt, dass es sinnvoll wäre, dass Regulierung und Gesetzgebung da ein bisschen flexibel reagieren könnten. Vielleicht an Sie die Frage, Frau

Wachter, wie ist das in der Praxis überhaupt möglich? Kann man in einem Jahr dann noch ein Addendum zum AI Act machen? Wie könnte man überhaupt gesetzgeberisch flexibel auf solche Sachen reagieren? Oder ist die EU dafür einfach ein viel zu langsamer Koloss?

Sandra Wachter [00:44:55]

Nein. Also diese Sache, dass das Recht immer so langsam ist, ist so eine Geschichte, das ist so ein Fake News, das immer herumgeistert. Recht kommt dann, wenn es Zeit ist, für [das] Recht zu kommen, nicht zu früh und nicht zu spät. Und es geht nicht darum, der Schnellste oder der Beste zu sein, sondern es geht darum, ein Recht zu schaffen, das vernünftig Risiken abdeckt. Das ist das, was man machen möchte. Somit halte ich das für kein Problem. In Wirklichkeit … sorry, was war die Frage noch einmal, ich habe es vergessen?

Moderator [00:45:31]

Entschuldigung?

Sandra Wachter [00:45:32]

Die Frage.

Moderator [00:45:33]

Ach so, die Frage, wie denn die Gesetzgebung flexibel reagieren könnte?

Sandra Wachter [00:45:36]

Ja, das kann man auf verschiedenste Art und Weise machen. Man kann zum Beispiel über die delegierten Rechtsakte die Risikogruppen noch einmal nachbessern. Das sollte kein Problem sein. Dann kann man mit den Standards, die jetzt auch noch geschaffen werden, in den nächsten ein, zwei Jahren konkretere Maßnahmen setzen, das ein bisschen ausarbeiten. Wir haben ja auch noch Zeit bis ins nächste Jahr Feintuning selbst vom AI Act zu machen. Es ist ja auch kein Problem, im AI Act selber gibt es auch [Revision Periods], die eingesetzt sind, dass man alle paar Jahre einfach schauen muss, wie effektiv das Regelwerk eigentlich ist, und dann kann man Nachbesserungen treffen. Wie gesagt, diese Idee, dass das Recht immer so hinten nachhängt und man nichts mehr machen kann, wenn es da ist, [es] sofort outdated ist, ist einfach nicht wahr. Eigentlich ist der [AI Act] relativ gut geschrieben, um flexibel auf solche Änderungen [reagieren] zu können. In Wirklichkeit, ja, ist es aus meiner Sicht kein Problem.

Moderator [00:46:35]

Herr Hacker, noch kurz. Ansonsten würde ich hier noch schnell mit den letzten Fragen weitermachen.

Philipp Hacker [00:46:40]

Ganz, ganz kurz. Und was jetzt spezifisch diese Thresholds anbelangt, diese Schwellenwerte, da sind genauso delegierte Rechtsakte der Kommission angedacht. Das kann man dann auch relativ spontan ändern, und da gibt es, wie Sandra sagt, ganz viele Techniken im Recht, eines zum Beispiel

auch mit Generalklausel. Alle beschweren sich immer: Oh, es ist alles so unbestimmt, die Rechtsbegriffe. Das ist Absicht, das ist strategische Unbestimmtheit, um neue Entwicklungen aufnehmen zu können. Und dann gibt es Standards, die auf die jetzigen Entwicklungen reagieren und die dann entsprechend schneller angepasst werden können. Die Rechtswissenschaftler haben ja mit dem Problem auch nicht seit gestern erst zu tun, sondern [seitdem] es Technologien gibt, ist das ein Dauerbrenner. Das ist so eine Urban Legend, dass das Recht zu spät kommt.

Moderator [00:47:24]

Okay, dann noch direkt an Sie, Herr Hacker, die Nachfrage, die es hier auch noch im Chat gab, ob von militärischen KI-Projekten etwas bekannt ist, die da auch mitagieren, und ob es nicht ein großes Sicherheitsrisiko wäre, wenn die eigene Verteidigung und Sicherheit auf ausländischen Modellen basiert, von denen man dann natürlich abhängig ist und die vielleicht auch potenzielle Backdoors haben?

Philipp Hacker [00:47:47]

Ja, in der Tat. Es gibt, das ist öffentlich verfügbare Information, das erste KI-Rüstungs-Unicorn, das sind ja Start-ups, die mit mehr als einer Milliarde bewertet werden. Das ist Helsing, europaweit das erste, drei Gründer, die sind auch hier in Berlin und in München ansässig. Das sind wirklich hochprofessionelle Leute, die auch entsprechend militärische Erfahrung mitbringen und die da wirklich hervorragende Arbeit leisten. Es gibt natürlich im Ausland, auch vor allem in den USA, einige Start-ups, die da sehr aktiv sind, Shield AI zum Beispiel. Aber da geht es natürlich ganz zentral darum, Modelle zu verwenden, bei denen wir uns darauf verlassen können, dass die entweder von Deutschland selbst oder von ganz engen NATO-Partnern gebaut werden. Selbst da muss man sagen, müsste man schon vorsichtig sein, wenn man ein US-Unternehmen nutzt. Da muss man sich schon sehr klar vertraglich absichern, dass da seitens der Regierung nicht interveniert werden kann.

Deshalb ist es so wichtig, diese Capabilities hier auszubauen. Da haben wir tatsächlich ein richtig gutes Unternehmen in Deutschland, das auch schon sehr etabliert ist und ich würde sagen europaweit auch seinesgleichen sucht und wirklich mit in der Weltspitze mitmischt. Und das ist auch extrem wichtig. Denn wir sehen ganz klar, die Strategie in China und auch in Russland ist, die teilweise noch defizitären Waffensysteme, was jetzt die eigentliche Waffenproduktion anbelangt und die Genauigkeit der Waffensysteme – sieht man in der Ukraine sehr gut, dass Russland zum Beispiel schwer unterlegen ist den NATO-Standards –, diese Defizite wettzumachen durch technologische Mittel. Und dem muss man sich einfach stellen.

Wir leben jetzt in einer Welt, in der leider diese Bedrohungen allenthalben gängig sind, wenn wir uns die Achse anschauen. Auch Iran verfügt über sehr gute technische Möglichkeiten, und da muss man einfach mithalten und muss auch in dem Bereich die Scheuklappen ablegen. Zu Recht haben wir in Deutschland sehr lange gesagt, wir trennen ganz stark Ziviles und Militär, auch gerade in der Forschung. Das muss man neu überdenken im Rahmen der Zeitenwende. Wir brauchen auch eine Art von Zeitenwende in der KI.

Moderator [00:50:04]

Vielen Dank. Und dann zum Schluss habe ich noch eine Frage an Sie, Frau Wachter. Ist jetzt vielleicht ein bisschen unfair, weil ich eigentlich gedacht hatte, die Verhandlungen sind zu diesem Zeitpunkt schon vorbei. Es scheint sich jetzt nicht anzudeuten, wie mir eben noch gesagt wurde. Wie geht es denn jetzt eigentlich weiter mit dem AI Act? Kann man das schon sehen? Oder hängt das von den nächsten zwei Stunden im Trilog ab?

Sandra Wachter [00:50:27]

Es hängt jetzt einmal davon ab, auf was man sich grob einigt. Ich weiß nicht, wird das heute oder morgen sein, aber dann ist die Arbeit ja immer noch nicht getan. Wie gesagt, das Feintuning kommt ja dann auch im nächsten Jahr. Dann hat man ja auch noch die Aufgabe, die Standards zu schaffen, die dann auch noch gemacht werden. Es wird auch noch ein, zwei Jahre dauern wahrscheinlich, bis die ausgefertigt [sind], weil man hat diesen AI Act, der das Grobe angibt, was gemacht werden soll. Und die Spezifizierungen sind dann in diesen sogenannten Standards festgelegt. Das wird auch noch dauern, bis die hergestellt werden. Deswegen ist das auch noch nicht fertig. Und selbst wenn das Ganze jetzt verabschiedet wird, dann hat man ja noch eine Übergangszeit, die zwischen zwei und fünf Jahren [dauern] kann. Man darf den Atem gespannt halten. Wie das im Endeffekt aussehen wird, wird sich jetzt in den nächsten Tagen, Wochen, wahrscheinlich Monaten erst zeigen.

Moderator [00:51:23]

Ja, Herr Hacker, war das eine Meldung?

Philipp Hacker [00:51:26]

Nur ganz kurz. Couldn't agree more. Und eine Sache, die auch noch spannend ist zu sehen, die AI Liability, also die Schiene Product Liability Update, also Produkthaftungsrichtlinie, und eventuell auch die AI Liability Directive. Auch das ist etwas, was dann noch auf den AI Act abgestimmt wird. Und zumindest die Überarbeitung der Produkthaftungsrichtlinie, die wird wahrscheinlich gleichzeitig mit dem AI Act kommen und das ist eminent wichtig für viele Unternehmen, vielleicht sogar noch wichtiger als der AI Act. Denn da wird nicht nur AI, sondern auch Software erfasst. Und da geht es ganz eminent auch um Schadensersatzklagen, die gerade kleinere Unternehmen sehr schnell in den Ruin treiben können.

Moderator [00:52:02]

Ja, Frau Wachter.

Sandra Wachter [00:52:03]

Und zu dem Punkt auch noch: Und da tut es mir eigentlich in der Seele richtig leid, dass die Medienberichterstattung sich für diese beiden Regulierungen nur sehr stiefkindlich interessiert, weil es eben darum geht, dass es da Haftungsmöglichkeiten für Individuelle gibt. [Da müssten wir eine ganz andere] Pressekonferenz haben, ob das gut ausgestaltet ist und so weiter. Das ist eine Regulierung, die relativ wenig Sonnenlicht bekommt. Und es wäre eigentlich besser, da auch mehr Public Discourse darüber zu haben, um das auch noch mal verbessern zu können.

Moderator [00:52:37]

Perfekt. Vielen Dank. Herr Hacker, was Sie da gerade geschrieben haben, das sehen die im Chat nicht. Sie müssen das noch einmal unten einstellen, damit Sie es an alle schicken. Ah, hat unser Tech Support schon gemacht, hervorragend. Dann ist die Zeit jetzt auch vorbei. Erst einmal vielen Dank an die Journalistinnen und Journalisten, die dabei waren. Vor allem natürlich vielen Dank an Sie, Frau Wachter und Herr Hacker, dass Sie dabei waren. Wir werden heute so schnell wie möglich die Aufzeichnung von der Veranstaltung auf unserer Homepage online stellen. Voraussichtlich

morgen gibt es dann das redigierte Transkript. Falls Sie die Audioaufzeichnungen, die Videodatei oder heute schon das maschinell erstellte Transkript haben wollen, dann finden Sie in der Reminder-Mail von heute Morgen, die Sie bekommen haben, den Link dazu. Vielen Dank für Ihre Zeit. Ich denke, heute werden Sie noch einiges mit Medienanfragen zu tun haben. Dann wünsche ich Ihnen noch einen schönen Tag und auf Wiedersehen. Vielen Dank!

Sandra Wachter [00:53:29]

Vielen Dank für die Einladung.

Philipp Hacker [00:53:31]

Vielen Dank.

Ansprechpartner in der Redaktion

Bastian Zimmermann
Redakteur 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