13.05.2026

Transkript

„Wie verändert künstliche Intelligenz Cybersicherheit?"

Expertinnen und Experten auf dem Podium

- **Dr. Jonas Geiping**
  Leiter der Forschungsgruppe „Safety- Efficiency- aligned Learning" am ELLIS Institut Tübingen und am Max-Planck-Institut für Intelligente Systeme, Tübingen
- **Prof. Dr. Thorsten Holz**
  wissenschaftlicher Direktor, Max-Planck-Institut für Sicherheit und Privatsphäre, Bochum
- **Prof. Dr. Konrad Rieck**
  Leiter des Lehrstuhls für Maschinelles Lernen und Sicherheit, Berlin Institute for the Foundations of Learning and Data (BIFOLD), Technische Universität Berlin
- **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/69fde509f79ef1af9f532ec0

Transkript

Moderatorin [00:00:00]

Herzlich willkommen, liebe Journalistinnen und Journalisten, hier zum Press Briefing des Science Media Center dazu, wie künstliche Intelligenz die Cybersicherheit verändert. Mein Name ist Samantha Hofmann. Ich bin Redakteurin für Digitales und Technologie hier beim Science Media Center und mit mir heute hier sind drei Experten. Deren Expertise reicht von künstlicher Intelligenz (KI) bis Cybersicherheit, alle mit unterschiedlichen Schwerpunkten. Ich stelle sie dann gleich auch noch genauer vor, aber erst noch einige organisatorische Sachen. Liebe Journalistinnen und Journalisten, Ihre Fragen stellen Sie bitte über die F von Zoom. Sie posten Ihre Fragen da rein. Ein Kollege von mir sammelt sie und ich stelle sie dann hier im Gespräch an die Experten. Nutzen Sie bitte ausschließlich das F für Ihre Fragen und nicht den Zoom-Chat. Das macht es für uns einfach sehr viel übersichtlicher. Außerdem besteht in dem F die Möglichkeit, Fragen mit einem Daumen nach oben zu bewerten. Machen Sie das bitte auch gerne viel. So können wir sehen, welche Fragen für Sie besonders relevant sind und die stellen wir dann gegebenenfalls priorisiert, sollte es zu viele Fragen geben. Sie müssen nichts mitschneiden, was hier passiert. Wir zeichnen das Meeting auf und stellen ein Transkript sowie eine Audio-und Videodatei nach dem Meeting zur Verfügung. Jetzt zum Thema: Dass KI die Cybersicherheit verändert, ist mittlerweile klar. Allerdings ist immer noch die Frage: Wie genau und wie stark? Glaubt man den großen Anbietern von KI-Modellen, dann sehr stark. Aus dem Grund haben zum Beispiel Anthropic und OpenAI Varianten ihrer neuesten Modelle nicht öffentlich zugänglich gemacht. Die Sprachmodelle sollen einfach zu gut darin sein, Schwachstellen in Code zu finden. Allerdings ist es ein Unterschied, ob man eine Schwachstelle findet oder ob man sie auch ausnutzen kann. Wie gut KI-Modelle im Ausnutzen von Schwachstellen sind, haben Thorsten Holz und sein Team sich angeschaut. Dazu haben sie unter anderem auch mit Anthropik, OpenAI und Google zusammengearbeitet. Thorsten Holz wird uns gleich noch Genaueres dazu erzählen, was sie gemacht haben und was für Ergebnisse dabei rausgekommen sind. Aber erst stelle ich Ihnen unsere Experten heute einmal vor. Ich beginne mit Jonas Geiping. Er leitet die Forschungsgruppe „Safety- Efficiency- Aligned Learning" am ELLIS Institut in Tübingen und er beantwortet heute unsere Fragen dazu, wie KI und Sprachmodelle sicher gemacht werden können, sodass sie im besten Fall gar nicht dabei helfen, Cyberangriffe durchzuführen. Dann eben schon erwähnt, Thorsten Holz. Er ist wissenschaftlicher Direktor am Max-Planck-Institut für Sicherheit und Privatsphäre in Bochum und er beantwortet Fragen dazu, wie IT-Infrastruktur sicher gemacht werden kann und welche Rolle KI für Cybersicherheit spielt. Die Runde abschließt Konrad Rieck. Er leitet den Lehrstuhl für Maschinelles Lernen und Sicherheit am Berlin Institute for the Foundations of Learning and Data an der TU Berlin. Er verbindet ein bisschen die Bereiche KI und Cybersicherheit und weiß, zu welchen Cyberangriffen KI-Systeme in der Lage sind und wie man sich gegen diese Angriffe schützen kann. Ich habe für Sie alle drei eine Eingangsfrage vorbereitet und würde gerne mit Ihnen beginnen, Herr Rieck. Und zwar mit der Frage: Wie verändert KI den Wettstreit zwischen denen, die IT-Schwachstellen ausnutzen und denen, die IT-Infrastruktur schützen möchten? Hilft sie dabei eher den Angreifern oder den Verteidigern?

Konrad Rieck [00:03:07]

Wenn man die Frage so stellt, muss man ein bisschen gucken, was es da für Ebenen gibt. Ich glaube, wir sind in einer sehr spannenden Zeit. Es passiert sehr viel in der IT-Sicherheit, aber nicht alles, was auf den ersten Blick den Angreifern hilft, nützt den Verteidigern gar nichts. Zum Beispiel bin ich der Meinung, das Finden von Schwachstellen ist eigentlich immer etwas Gutes. Auch wenn wir das jetzt so hören, dass das ganz schlecht ist. Aber eigentlich müssen wir Schwachstellen finden und sie entfernen. Sie sind da. Es ist nicht so, dass die KI die Schwachstellen in die Software macht, sondern wir haben diese unsichere Software und die läuft jeden Tag. Und eigentlich ist jede Schwachstelle, die wir finden, potenziell eine Schwachstelle weniger. Ist ein bisschen die Frage,

wer sie findet. Aber theoretisch sind solche Techniken erst mal sehr gut. Was jetzt ein bisschen Unruhe in die Sache bringt, ist, dass die modernen Modelle, so was wie Claude Mythos, eben nicht nur Schwachstellen finden, sondern dass sie auch gut sind, Schwachstellen auszunutzen. Das ist was, was man in der IT-Sicherheit auf den ersten Blick nicht braucht, wenn man Verteidiger ist. Und das spielt den Angreifern in die Hände. Gleichzeitig ist es aber auch so: Damit eine Schwachstelle eine Schwachstelle ist, muss man sie auch ausnutzen können. Das heißt, wenn ein Modell sagt, da ist was kaputt, das ist ein Sicherheitsproblem, aber das Modell kann dir gar nicht sagen, warum eigentlich, dann ist das nicht so gut. Es ist eigentlich auch besser für Verteidiger, wenn man tatsächlich weiß, dass eine Schwachstelle auch wirklich eine Schwachstelle ist. Was ich so kompliziert hier sagen will, ist, ich glaube, diese Technologie hebt die Automatisierung in der IT-Sicherheit an. Das heißt, es werden schneller Schwachstellen gefunden, schneller ausgenutzt, wahrscheinlich auch schneller gepatched. Und aus meiner Sicht würde ich sagen, zurzeit gibt es einen Vorteil für die Angreifenden. Aber ich sehe den nicht so riesig, dass wir uns jetzt komplett umstellen müssen. Sondern wir müssen diese Welle von neuen Schwachstellen quasi überleben und danach werden die Systeme wahrscheinlich sicherer werden.

Moderatorin [00:05:00]

Und „gepatched" bedeutet, ich finde eine Schwachstelle und repariere sie dann quasi.

Konrad Rieck [00:05:04]

Korrekt.

Moderatorin [00:05:06]

Genau. Okay, vielen Dank. Dann meine nächste Frage an Thorsten Holz, weil wir es eben schon kurz angesprochen haben: das Ausnutzen von Schwachstellen. Darum geht es in Ihrem Paper, das Sie jetzt hochgeladen haben. Können Sie kurz erläutern, was genau Sie da gemacht haben und was die wichtigsten Erkenntnisse daraus sind?

Thorsten Holz [00:05:23]

Genau. Wir haben in Zusammenarbeit mit Forschenden von der UC Berkeley und von der Arizona State University und einigen Industriepartnern, also allen Frontier Labs von Anthropic, OpenAI und auch Google, einen Benchmark entwickeln. Das Ziel ist, in einer kontrollierten Umgebung besser zu verstehen, wie die Fähigkeiten dieser Modelle sind. Vor einigen Wochen hatte der Blogpost von Anthropic zu Mythos für sehr viel Diskussion gesorgt. Wobei, das muss man natürlich auch immer mit einer gewissen Marketingbrille sehen. Das ist eine Firma, die jetzt kurz vor dem Börsengang steht, die natürlich auch Aufmerksamkeit erregen will. Deshalb war das Ziel unseres Benchmarks, einfach mal in kontrollierten Umgebungen zu überprüfen: Wie gut sind diese Modelle eigentlich? Nicht nur im Finden von Schwachstellen, sondern insbesondere in der Ausnutzung, also der Erzeugung eines konkreten Exploits, der dann demonstriert, dass eine Schwachstelle auch wirklich einen Security Impact hat, wie Herr Rieck auch gerade erläutert hat. Was wir dazu gemacht haben, ist, wir haben etwa neunhundert verschiedene Challenges gebaut. Das sind teilweise Programme, die einfach normale Nutzerprogramme sind. Allerdings haben wir auch den Chrome Browser und den Linux Kernel eingebaut, also zwei sehr, sehr schwierige Ziele als mögliche Challenges. Um das einzuordnen: Für einen Menschen dauert eine Ausnutzung einer solchen Schwachstelle im Browser beziehungsweise im Kernel mehrere Tage oder vielleicht auch mehrere Wochen. Und das Ziel war zu testen, wie gut ist ein Modell jetzt in der Ausnutzung davon. Was wir dann sehen, ist, dass vor allem Claude Mythos beziehungsweise GPT-5.5, also die beiden aktuellen Frontier Modelle,

erstaunlich gut in dieser Aufgabe sind. Sie schaffen es in einer Vielzahl von Fällen, bei Mythos waren es so 160, bei GPT so 120 Fälle, anhand von einem gegebenen Crash, wirklich ein Ende-zu-Ende-Exploit zu erstellen. Das Ganze innerhalb von zwei Stunden also und auch komplett automatisiert. Der Lichtblick ist vielleicht ein bisschen, dass, sobald wir dann eben noch verschiedene Arten von Schutzmechanismen anschalten, haben die Modelle es schwieriger. Wir sehen also, dass sie damit noch nicht komplett klarkommen. Allerdings muss man das auch im Kontext sehen. Das sieht man auch, wenn man ältere Modelle benutzt, sind die gar nicht in der Lage, überhaupt eine dieser Schwachstellen auszunutzen. Und wir haben jetzt doch sehr schnell sehr große Fortschritte in letzter Zeit beobachten können.

Moderatorin [00:07:48]

Sie haben jetzt eben von 160 beziehungsweise 120 Fällen gesprochen, in denen Schwachstellen ausgenutzt werden konnten. Wie viele Fälle wurden denn insgesamt betrachtet?

Thorsten Holz [00:07:58]

So etwas mehr als 900.

Moderatorin [00:08:01]

Okay. Also doch schon eine jetzt keine übermäßig große Zahl, aber doch eine relevante Zahl an Fällen, die ausgenutzt werden können.

Thorsten Holz [00:08:09]

Genau.

Moderatorin [00:08:10]

Okay. Vielen Dank. Noch eine ganz kurze Nachfrage. Die Modelle, die Sie da genutzt haben, das Modell von Anthropic, das Sie angesprochen haben, ist das, was noch nicht öffentlich zur Verfügung steht, richtig?

Thorsten Holz [00:08:20]

Genau.

Moderatorin [00:08:21]

Das von Open AI, ist das das GPT-5.5-Cyber, das noch nicht zur Verfügung steht genau oder ist es das öffentliche?

Thorsten Holz [00:08:29]

Nee, das sind jeweils die, also Mythos Preview und eben GPT-5.5-Cyber. Also wir haben den Benchmark entwickelt, haben den Benchmark dann den Kollegen von Anthropic beziehungsweise OpenAI gegeben. Die haben die Auswertung gemacht. Wir haben auch die Resultate gesehen und

auch dann die Analyse durchgeführt. Also wir haben keinen direkten Zugriff auf die Modelle, allerdings haben die Teams von dort auf unserem Benchmark die Evaluierung durchgeführt.

Moderatorin [00:08:58]

Alles klar. Dann gerne die Frage an Herr Geiping. Und zwar, wie versuchen denn Anbieter sicherzustellen, dass Sprachmodelle eben nicht für diese „bösen" Zwecke der Cyberangriffe genutzt werden können? Und wie zuverlässig sind solche Verfahren allgemein, jetzt vielleicht ohne schon über die Details zu sprechen, wie genau diese Methoden aussehen? Einfach einen Überblick darüber.

Jonas Geiping [00:09:21]

Genau, weil es keine einfache Art und Weise gibt, diese Angriffe einfach abzuschalten, gibt es oft so eine Art von Defence in Depth in den Modellen. Zum Beispiel dadurch, dass die Firmen versuchen, mit dem Klassifikator Anfragen zu detektieren, die zum Beispiel so aussehen, als wären sie ein Angriff auf eine andere Codebase oder ein Angriff auf jemand anders. Das wäre vielleicht die erste Schicht, dass vielleicht dieser Klassifikator bemerkt, dass ein Angriff stattfindet und dass diese Eingabe nicht durchgehen sollte. Dann eine zweite Schicht wäre, dass das Modell darauf trainiert wird, dass es erkennen kann, das ist hier eine Situation, wo wirklich jemand versucht, eine Codebase zu brechen und das wäre ein Schaden für eine andere Partei, wenn dieser Anruf stattfindet. Dass das Modell das durch sein Training in solchen hypothetischen Fällen erkennt und sagt, jetzt machen wir hier nicht weiter. Das ist natürlich alles gar nicht so einfach, weil oft die Firmen hier ein bisschen im Hintertreffen sind. Sobald die Firmen zum Beispiel gesehen haben, wie Angriffe durchgeführt werden mit den Modellen, dann können sie in der nächsten Schicht oder in der nächsten Phase die Modelle drauf trainieren, dass solche Angriffe nicht mehr funktionieren. Zum Beispiel ein klassischeres Problem wäre, dass man älteren Modellen noch oft sagen konnte, wir machen eine Challenge und wir testen nur Capture the Flag und das ist alles hypothetisch und nichts passiert wirklich. Und die Modelle sagen, dann, spielen wir Capture the Flag zusammen und wir testen diesen Angriff und wir greifen dieses hypothetische System an. Das ist auch was, was bei den besseren Modellen jetzt mittlerweile nicht mehr funktioniert. Aber weil das alles so ein bisschen retroaktiv ist, ist es hier oft schwierig für die Firmen, wirklich zu 100 Prozent sicherzustellen, dass nichts passiert. Das andere Problem ist, dass die Modelle oft einen fehlenden Kontext haben. Was man oft machen kann, ist so eine Dekompositionsattacke, in der das Ausnutzen einer Schwachstelle auf mehrere Unterbereiche verteilt wird. Zum Beispiel verschiedene Eingaben, die vollkommen separat voneinander sind. Da wird dann versucht, dass Teile dieses Angriffs voneinander getrennt werden. Dass nur der Benutzer, der diesen Angriff durchführt, überhaupt alle Teile sieht und das dann zusammensteckt. Und das Modell selber nur einen Teil sieht. Zum Beispiel, wir suchen einen Bug in diesem Teil von Firefox und ein anderer Kontext von dem Modell sieht, wir suchen einen Bug in dem Teil von Firefox und wir versuchen, das zu patchen. Und dass dann von außerhalb das alles zusammengesetzt wird. Und das ist natürlich sehr schwer für das Modell zu verstehen, dass hier ein Angriff stattfindet, weil da viel Kontext fehlt.

Moderatorin [00:12:08]

Eine Möglichkeit, Modelle dann doch zu Cyberangriffen zu bringen, ist, sie so ein bisschen zu überlisten, ihnen den Kontext wegzunehmen, zu sagen, du weißt überhaupt nicht, was ich hier mache und dann die einzelnen Informationen selbstständig zusammenzutragen.

Jonas Geiping [00:12:22]

Genau, das wäre ein Beispiel für eine Attacke, die oft auch bei aktuellen Modellen noch funktioniert.

Moderatorin [00:12:25]

Alles klar, vielen Dank. Eine Frage an Sie, Herr Holz. Und zwar glaube ich, ist es noch ganz hilfreich, wenn wir uns erstmal angucken, was es überhaupt für Arten von Cyberangriffen gibt. Es gibt zum Beispiel Phishing-Kampagnen oder Programme, die gehackt werden, Ransomware-Attacken. Vielleicht können sie da einmal einen Überblick über die relevantesten Cyberangriffsmöglichkeiten geben und dann auch ganz kurz sagen, welche Rolle KI bei den verschiedenen Arten spielen kann.

Thorsten Holz [00:12:56]

Genau. In der Praxis gibt es natürlich jetzt nicht nur diese Softwareschwachstellen, auf die wir uns hier konzentriert haben, sondern viele verschiedene Arten von Angriffen, wie eine Firma oder eine Behörde kompromittiert wird. In der Praxis spielt Social Engineering eine sehr große Rolle, weil der Faktor Mensch einfach von den Angreifern häufig ins Ziel genommen wird. Das ist etwas, wobei KI wahrscheinlich auch unterstützen kann. Weil zum einen kann man nun viel einfacher Daten sammeln, automatisiert auswerten und dann wirklich personalisierte Social Engineering Angriffe durchführen...

Moderatorin [00:13:28]

Können sie vielleicht ganz kurz erklären, was Social Engineering ist?

Thorsten Holz [00:13:31]

Social Engineering ist die Idee, dass ich als Angreifer zum Beispiel eine Email so personalisiere, um sich das Vertrauen des Opfers zu erschleichen. Dass ich das Opfer dazu bringe, zum Beispiel auf einen Link zu klicken oder ein Programm auszuführen, indem ich zum Beispiel vorgebe, das ist eine Mail, die kommt jetzt vom Chef. Was in der Praxis auch gut funktioniert: Dass man zum Beispiel vorgibt, das ist jetzt ein Scan, der von einem Drucker erstellt wurde. Hier ist der Scan, öffne das mal. Oder: Das eine Mail aus der Personalabteilung, hier gibt es eine Änderung in einer Abrechnung. Bitte schaut in den Anhang rein. Und KI-Methoden werden dabei helfen, das viel besser zu personalisieren. Neben Text wird, denke ich, in der Zukunft auch Sprache eine größere Rolle spielen. Sodass man dann auch Telefonanrufe bekommt. Das sind in der Praxis Enkeltrick und Co, die dann personalisiert werden. Dass man auch die Stimme hat, die dann zu einer gewissen Person besser passt. Ich denke in dem ganzen Social Engineering und Phishing Bereich wird KI eine große Rolle spielen. Und der andere große Bereich sind die Softwareschwachstellen. Dass die Angreifer in der Lage sind, ein System zu kompromittieren, beliebigen Code auszuführen, Ransomware zu installieren. Und wo KI auch eine große Rolle spielt, ist im Finden und Ausnutzen von Schwachstellen. Und ich denke in beiden oder in vielen Bereichen können sowohl Angreifer als auch Verteidiger in Zukunft auf KI zurückgreifen. Und das wird die ganze Dynamik in diesem Umfeld deutlich erhöhen und das ganze auch viel schneller machen, als wir das bis jetzt kennen.

Moderatorin [00:15:10]

Sieht man das vielleicht auch schon in Zahlen, dass in den vergangenen Jahren die Phishing-Angriffe hochgegangen sind?

Thorsten Holz [00:15:18]

Dazu habe ich keine Informationen. Das weiß ich nicht.

Moderatorin [00:15:22]

Wissen sie da was, Herr Rieck? Wenn nicht, ist auch kein Problem.

Konrad Rieck [00:15:24]

Also wir haben da eine Studie, die im August rauskommt. Da haben wir quasi getestet, wie effektiv so personalisiertes Phishing ist mit Leuten und wir konnten die Klickrate, also das ist quasi die Häufigkeit, dass Leute auf so eine Phishing-Email klicken, mit personalisierten Angriffen fast verdreifachen. Das heißt nicht, dass die danach dann auch wirklich alle ihre persönlichen Daten eingeben, aber sozusagen ein Schritt war gemacht. Und man kann sagen, dass die aktuellen Modelle diese Arten von Social Engineering wirklich verbessern. Allerdings, das muss man dazu sagen, das können alle Modelle, die es jetzt schon gibt. Also das ist jetzt nichts wwofür wir quasi die neuesten Frontier-Modelle brauchen. Weil wir in Deutschland natürlich nicht alle wilden Angriffe machen können, die wir uns vielleicht interessieren, haben wir das alles lokal gemacht mit Open-Weight-Modellen und die Angriffe haben trotzdem funktioniert. Jetzt kann man spekulieren, dass das mit moderneren und besseren Modellen noch besser funktioniert, aber ich würde sagen, diese Bedrohung ist schon da und man kann sie auch schon quantisieren. Also sie ist quasi eine mindestens Verdoppelung der Klickraten bei Phishing-Emails.

Moderatorin [00:16:38]

Da habe ich noch eine passende Frage zu an Herr Geiping. Wir haben nämlich vor diesem Meeting schon per Mail Fragen von Journalist:Innen bekommen. Die stelle ich gleich. Aber vorher möchte ich noch einmal darauf hinweisen, dass Sie, liebe Journalistinnen und Journalisten, natürlich sehr, sehr gerne in das F auch Ihre Fragen posten können. Das ist jetzt die Gelegenheit, ihre Fragen an die Experten loszuwerden. So die Frage an Herr Geiping: Anthropic und OpenAI haben sich entschieden, Varianten ihrer neuen Modelle nicht zu veröffentlichen. Kann das eine dauerhafte Lösung sein? Wir haben gerade auch schon gehört, dass gar nicht nur die neuesten Modelle ein Problem werden können. Besteht dann die Gefahr, dass die Modelle doch noch an die Öffentlichkeit kommen oder ist das genau ausgeschlossen?

Jonas Geiping [00:17:20]

Also ausgeschlossen ist natürlich nie, dass das Modell irgendwie geleaked wird oder dass jemand Zugriff bekommt, der keinen Zugriff haben sollte. Ich glaube auch bei Anthropic gab es einen Vorfall, wo jemand, ich glaube ein Team in Singapur, Zugriff hatte auf das Modell, was eigentlich nicht war. Aber trotzdem, diese Limited-Release-Phase ist glaube ich schon ein guter Schritt. Eben weil dadurch dann auch Angriffe intern getestet werden können und es können Vulnerabilities gescannt werden. Und weil ich vorher gesagt habe, dass Firmen mit ihren Safeguards ein bisschen im Hintertreffen stehen: Durch diesen Limited Release können sie halt so schauen, was das Modell kann und mit gewissen Partnern gucken, was irgendwie auf der Modellseite gepatcht oder

detektiert werden muss. Und das kann den Firmen helfen, das Modell ein bisschen sicherer zu machen, wenn es dann an alle released wird. Also das, denke ich, ist schon hilfreich. So weit haben wir das gesehen von den Modellen von Anthropic und von OpenAI. Aber zum Beispiel die tiered Modelle, die sind zum Teil auch jetzt gut und werden noch besser. Und da ist dann auch die Frage, wie sich das hier ausspielt mit den verschiedenen Anbietern. Es gibt vielleicht auch den Test von Anbietern, ihr Modell möglichst schnell anzubieten, möglichst breit, auch wenn noch nicht vollständig getestet ist, was das Modell kann und wo das Modell Angriffe finden kann, die vielleicht vorher noch undenkbar waren.

Moderatorin [00:18:57]

Aber Sie teilen auch die Ansicht, die wir eben schon gehört haben, dass es gar nicht unbedingt die neuesten Modelle braucht, damit Cyberangriffe durchzuführen?

Jonas Geiping [00:19:05]

Für die persönlichen, also für für Phishing-Bereiche, wie wir vorhin gesprochen haben, da stimme ich auf jeden Fall zu. Aber jetzt grade im Bereich von Software Vulnerabilities haben wir schon gesehen, dass die neuesten Modelle besser sind. Das kommt auch im Report von Herr Holz vor, dass zum Beispiel GLM, ein Modell von vor vier oder fünf Monaten, noch fast keine einzige Vulnerability findet. Und das ist schon irgendwie spannend, dass es eine Frontier gibt, an der wirklich neue Schwachstellen gefunden werden, von den stärksten, größten, am meisten trainierten Modellen. Und das ist eigentlich eine ganz spannende Dynamik, weil hier plötzlich sich das Kraftverhältnis ein bisschen verschiebt. Also, es sieht fast schon ein bisschen so aus, als würde es alle Seiten mehr kosten, diese Schwachstellen zu patchen. Weil vielleicht in der Zukunft jede Firma das teuerste Modell laufen lassen muss für ein paar Monate, um im kommenden Zyklus sicher zu sein, wenn dann dieses teuerste Modell für alle zugänglich ist. So ein Wettstreit hier. Aber wir müssen mal schauen, wie sich das entwickelt.

Moderatorin [00:20:23]

Ich glaube, Herr Holz wollte da noch etwas zu ergänzen.

Thorsten Holz [00:20:27]

Genau, wir hatten uns auch einige offene Modelle angeschaut und dabei war nur GLM, also das Modell von Z.ai, in der Lage, überhaupt Schwachstellen automatisiert zu exploiten. Wohingegen Mistral oder andere offene Modelle kamen mit der Aufgabe gar nicht klar. Die waren nicht in der Lage, eine dieser Challenges zu lösen. Vielleicht um das einzuordnen: In unseren Benchmarks war GLM etwa so auf dem Level von Claude Opus 4.6 oder 4.7, also schon ein Stück hinter Mythos beziehungsweise GPT-5.5. Allerdings ist das jetzt auch nicht unbedingt absehbar, ob die nicht innerhalb von drei oder sechs Monaten auch auf einem ähnlichen Level sind. Weil wir sehen generell, dass die offenen Modelle schnell aufholen. Da wird sich also auch in nächster Zeit einiges tun. Und gerade, wenn man diesen Modellen dann noch mehr Tools zur Verfügung stellt oder generell einfach die ganze Automatisierung verbessert. Ich denke, dann werden auch offene Modelle bald in der Lage sein, ähnliche Aufgaben zu lösen. Und dann ist natürlich die Frage, wie man da eine Kontrolle hat. Weil die kann jeder prinzipiell aufsetzen und auch entsprechend nutzen. Da wird es vermutlich dann noch Überlegungen geben müssen, wie man da besser mit umgeht.

Moderatorin [00:21:43]

Open-Source-Modelle, also offene Modelle könnten das auch demnächst können. Was bedeutet denn demnächst? Haben Sie da einen ungefähren Zeithorizont, zwei Jahre, fünf Jahre, wenn man sich die Entwicklung so anschaut?

Thorsten Holz [00:21:54]

Ich hätte eher gesagt sechs Monate bis ein Jahr. Also wir sehen schon, dass die in vielen anderen Benchmarks auch hinter den von Frontier Labs sind, allerdings dann doch innerhalb von Monaten oder innerhalb von zwölf Monaten vielleicht aufholen.

Moderatorin [00:22:11]

Okay, ich glaube, wir hatten da Ergänzungen zu. Gerne kurz, weil wir noch weitere Fragen haben. Erst Herr Geiping und danach gerne Herr Rieck.

Jonas Geiping [00:22:18]

Genau, also mich hat da zum Beispiel Kimi 2.6 interessiert. Was jetzt auch ein neuerer Release ist. Das ist zum Beispiel bei uns in manchen Tests auch schon sehr viel besser als das GLM-Modell. Vielleicht ist bei diesen Modellen sechs Monate schon eine lange Zeit. Es könnten auch vielleicht eher vier bis sechs Monate sein. Und was da auch spannend ist, vielleicht auch als Nachtrag zu dem, was Thorsten schon gesagt hat, was es auch bisschen schwierig macht, das abzuschätzen, ist auch diese Scaffold-Frage. Also es kann sein, dass es eine gewisse Art von Overhang gibt. Dass die Open-Source-Modelle das schon können oder dass auch Modelle generell eine Aufgabe schon können, aber dass noch niemand den richtigen Block drumherum gebaut hat, wie das Modell am besten genutzt werden kann, um diesen Angriff durchzuführen. Also es kann zum Beispiel sein, dass manche Angriffe schon funktionieren, aber es muss das richtige Tooling gebaut werden für das Modell. Und das ist alles schwer für uns alle vorherzusagen, weil eben das Tooling könnte irgendwie clever sein und irgendwie unerwartet und das Modell unerwartet plötzlich stark verbessern.

Moderatorin [00:23:15]

Bedeutet, es braucht gar nicht unbedingt eine langwierige Entwicklung, die da hinführt, sondern es reicht jetzt eigentlich der eine Moment, in dem jemand eine gute Idee hat und dann funktioniert das.

Jonas Geiping [00:23:24]

Das kann zum Teil für manche Modelle noch funktionieren, gerade für die Open-Source-Modelle, für die weniger Tooling passiert generell.

Moderatorin [00:23:30]

Herr Rieck, Sie wollten auch noch was ergänzen?

Konrad Rieck [00:23:33]

Was ein bisschen die Verbindung zu meinem Eingangsstatement herstellt: Also es gibt das Narrativ, dass, wenn Open-Weight-Modelle das alles können und das klang grad auch so ein bisschen an, dann kann jetzt jeder ein Hacker sein. Gleichzeitig, und das wird nämlich vergessen, kann aber auch jede Firma mit diesen Open-Weight-Modellen ihre eigene Software verbessern und diese ganzen Fehler finden. Und wir müssen so ein bisschen raus aus dieser Blase, dass wir immer denken, wenn wir so eine Technologie haben, dann wird die nur von den bösen Menschen benutzt. Ich würde vielleicht sogar sagen, wir müssen ein bisschen mehr Druck auf die Leute ausüben, die eben Software herstellen und sagen, ihr könnt euch jetzt nicht dahinter verstecken und sagen, es muss alles reglementiert sein. Sondern ihr müsst die neueste Technologie nutzen, um eure Software abzusichern. Ich will damit nicht sagen, dass die Welt dann automatisch sicher ist, aber es ist nicht so einseitig, wie es manchmal scheint.

Moderatorin [00:24:27]

Okay, das spielt schon auf eine Frage an, die wir auch im Vorfeld schon bekommen haben. Und zwar, seit Sprachmodelle es jedem ermöglichen, zu Vibecoden, Webseiten, Apps und verschiedene Anwendungen zu erstellen, bewegen wir uns auf eine Welt zu, in der Cybersicherheit kaum mehr garantiert werden kann? Das scheint mir jetzt so, als ob das nicht so wäre. Herr Rieck, gerne.

Konrad Rieck [00:24:48]

Also das ist eine sehr spannende Frage. Also wenn wir mal vordenken auf die nächsten fünf oder zehn Jahre, könnte es sein, dass wir in einer Welt leben, in der alle Software oder 99 Prozent der Software durch KI-Systeme erzeugt wird. Und ich kann aus meiner Forschung nicht sagen, ob das jetzt gut oder schlecht ist. Also schlecht wäre, dass der Mensch keine Kontrolle mehr hat und dass da viele Fehler drin sein können. Gut könnte aber auch sein, dass wir KI besser kontrollieren können als vielleicht Menschen. Also wir können die Fehler in den KI-Modellen beheben, wir können die Modelle besser machen. Was dann in fünf oder zehn Jahren passiert, weiß ich nicht, aber es wird sehr spannend.

Moderatorin [00:25:33]

Heißt, es kann passieren, dass KI durchaus mehr Sicherheitslücken einbaut in Code, den KI eben geschrieben hat?

Konrad Rieck [00:25:41]

Das ist schon so. Also ich glaube, da gibt es auch schon Studien.

Moderatorin [00:25:46]

Herr Geiping, Sie wollten dazu was sagen?

Jonas Geiping [00:25:48]

Das denke ich auch so. Also hier muss man vielleicht unterscheiden zwischen aktuellen Modellen, die zum Beispiel, wenn sie vollkommen automatisiert Webseiten oder Tools bauen, oft unsicherer sind als Experten, die so was bauen. Und Modellen in zwei, drei Jahren, für die das nicht mehr gilt.

Was natürlich auch weiterhin in vielleicht ein paar Jahren passieren kann, ist, dass es hier immer ein gewisses Machtungleichgewicht gibt, je nachdem, wie viel Geld oder wie viel Compute investiert wurde. Also zum Beispiel kann es sein, dass in zwei Jahren, wenn deine Website von der KI gebaut wird, die nicht sicher genug ist, wenn sie zu günstig war. Weil eben mehr Zyklen an Rechenleistung gebraucht würden, um sich gegen einen Angreifer zu verteidigen, der mehr Rechenleistung hat. Also das könnte sein, dass das weiterhin unsicher bleibt, je nachdem, wie viel Geld investiert wird. Das klang auch vorher an mit der Frage, dass auch die Firmen da mehr investieren müssen. Es kann sein, dass da mehr Kosten auf alle zukommen, um hier in diesem Wettstreit weiter mit gewissen Angreifern mitzuhalten.

Moderatorin [00:26:56]

Gerne dazu eine kleine Nachfrage an Herr Holz. Und zwar, wie sollen denn Nutzer, also zum Beispiel Unternehmen, auf diese Entwicklung reagieren? Ich kann mir vorstellen, dass auch gerade kleinere Unternehmen sich das nicht leisten können.

Thorsten Holz [00:27:09]

Also wir haben Kontakt mit ein paar größeren Firmen, die jetzt aktiv diese Modelle nutzen, um ihre Software intern zu testen. Vielleicht ein bekanntes öffentliches Beispiel ist Mozilla, die auch mit Anthropic zusammengearbeitet haben. Sie haben auch Blogposts dazu gemacht. Mythos hatte bei denen etwa 270 verschiedene Schwachstellen gefunden, die dann auch jetzt in den letzten Wochen gepatcht wurden. Also insofern, die großen Firmen verfügen, glaube ich, schon über die Fähigkeiten und viele, die ich kenne, setzen sich grade auch schon aktiv damit auseinander, um einfach jetzt schon mal proaktiv in ihrer Software nach Lücken zu suchen. Bei kleineren Firmen ist natürlich dann die Frage, ob die über die technische Kompetenz verfügen, wie sie am besten damit umgehen. Und vor allem auch die Kostenfaktoren, weil man muss auch sagen, diese Modelle sind nicht unbedingt günstig. Und vor allem, vielleicht auch ein interessanter Einblick aus unserem Benchmark war, dass die unterschiedlichen Modelle auch unterschiedliche Dinge finden. Also, wenn man sich dann so ein Venn-Diagramm, also eine Schnittmenge, erstellt von den Fähigkeiten der verschiedenen Modelle, sieht man, dass jedes Modell seine eigenen Fähigkeiten hat und es keine komplette Überlappung gibt. Und wir haben auch aktuell noch keine guten Metriken, um besser zu verstehen: Wie gut sind denn jetzt überhaupt diese Modelle? Was finden sie? Was können sie nicht finden? Ich glaube, da müssen wir generell noch mehr Forschung machen, um die Fähigkeiten besser zu verstehen und das dann eben auch so umzusetzen, dass das auch für Firmen besser nutzbar ist. Und vielleicht als erster Schritt in so eine Richtung haben sowohl OpenAI als auch Anthropic entsprechende Modelle oder entsprechendes Tooling zur Verfügung gestellt. Also Codex Security zum Beispiel. Da ist quasi die Idee, man gibt dem Modell Zugriff auf seinen Source Code, dann wird eine Überprüfung durchgeführt und das Ergebnis ist dann so eine Auflistung von verschiedenen Schwachstellen, die man gefunden hat. Also dass man eben auch die Nutzung von den Modellen einfacher macht und sich dann nicht etwas selber bauen muss, sondern es schon existierende Tools gibt. Und ich glaube, da wird es in Zukunft auch viel mehr Bedarf geben, wo dann auch noch viel mehr neue Lösungen entstehen müssen, um die Modelle auch einfacher nutzbar zu machen.

Moderatorin [00:29:14]

Wenn Sie jetzt sich die Cybersicherheit in Deutschland angucken mit den ganzen Schwachstellen, die es gibt, vielleicht wie die Systeme aufgebaut sind, könnten Sie dem deutschen Cybersicherheitssystem eine Schulnote geben? Und wieso? Gerne, Herr Holz.

Thorsten Holz [00:29:34]

Eine Schulnote. Ich denke, also das BSI koordiniert grade auch sehr sehr viel. Die Präsidentin hat vor ein paar Tagen auf LinkedIn beziehungsweise auf der BSI-Webseite auch ein größeres Posting abgesetzt mit verschiedenen Forderungen. Unter anderem, dass ein breiterer Zugriff auf die Modelle verfügbar sein sollte, dass man das besser koordiniert, dass man auch ein AI-Cyber-Security-Institut gründet, um sich dediziert mit diesen Fragen zu beschäftigen. Also ich denke, es gibt grade einige Aktivitäten und Initiativen, um mit den Frontier Labs proaktiver zusammenzuarbeiten. Und es gab da auch einigen Austausch, sowohl auf deutscher als auch auf europäischer Ebene. Ich denke, da gibt es gute Vorbereitungen. Aber ich denke, da sollte noch mehr getan werden, grade um einfach das Ganze etwas koordiniert zu machen oder auch, um Tools zu entwickeln, um die Modelle besser nutzbar zu machen. Also ich denke, auf einer eins sind wir noch nicht, allerdings auch nicht auf fünf. Ich würd dann vielleicht mal eine Zweiminus geben, weil einfach doch noch viel getan werden muss und generell die Investitionen in Sicherheit einfach noch hochgefahren werden müssen, um einfach unsere ganzen digitalen Infrastrukturen abzusichern.

Moderatorin [00:30:52]

Das klingt aber erstmal schon relativ solide. Wir haben von den Journalist:Innen noch eine Frage reinbekommen, die etwas grundsätzlicher ist. Und zwar fragt sich noch jemand, wie überhaupt Schwachstellen in Software entstehen können, also ob das alles Fehler sind, die irgendwann mal beim Programmieren gemacht wurden oder wie es dazu kommt. Vielleicht Herr Rieck.

Konrad Rieck [00:31:12]

Ja, also alle diese Fehler, die wir zurzeit finden, sind Fehler, die Menschen gemacht haben. Und es ist so, dass die Entwicklung von Software erstaunlich schwierig ist. Und es gibt schon Studien ich glaube aus den 60er Jahren, dass die NASA geprüft hat in der Apollo Mission, wie viel Zeilen falsch sind. Und ich glaube von 1000 Programmierzeilen waren drei falsch in der Apollo Software. Sie können sich vorstellen, dass das jetzt für irgendeine Webseite oder irgend so einen Shop im Internet natürlich noch viel mehr sind. Das heißt, es sind menschliche Fehler und einige von diesen Fehlern haben was mit Sicherheit zu tun. Viele aber nicht. Das muss man auch sagen. Also nicht jeder Bug, den zum Beispiel Mozilla reportet, der ist gleich ein Sicherheitsproblem. Aber je mehr eine Software im Kontext von Sicherheit verwendet wird, also wenn da Kreditkartendaten eingelesen werden oder andere persönliche Informationen, dann kann natürlich jeder Bug zu einer Sicherheitsschwachstelle werden. Und genau das habe ich am Anfang gesagt und das will ich auch noch mal wiederholen, weil das ist ein philosophisch interessanter Gedanke: Die Schwachstellen sind alle schon da, wir wissen nur nicht, wo sie sind. Wenn jetzt das nächste Modell noch mal 500 neue Schwachstellen findet und wir alle einen riesen Schock kriegen, die waren heute auch schon da. Und vielleicht gibt es auch Menschen oder Gruppen auf der Welt, die schon mehr Schwachstellen kennen, in Geheimdiensten, vielleicht auch unabhängig von KI. Das finde ich ja auch mal wichtig zu sagen. Es gab in der IT-Sicherheitsforschung eine ganze Reihe von Entwicklungen in den vergangenen zehn Jahren. Fuzzing zum Beispiel, was Herr Holz ja auch super viel gemacht hat.

Moderatorin [00:32:48]

Sorry, können sie ganz kurz den Begriff Fuzzing erklären?

Konrad Rieck [00:32:49]

Achso, Fuzzing ist eine Technik, um durch testen Fehler in Software zu finden. Man gibt einfach zufällige oder so semi-zufällige Inputs der Software und guckt, ob sie sich irgendwann irgendwie verschluckt. Und da gibt es Projekte von Google, die seit fast zehn Jahren laufen und die haben tausende von Fehlern gefunden. Diese Fehler sind halt überall und wir müssen halt nach und nach versuchen, so viel wie möglich zu beheben und dabei nicht aber weitere erzeugen. Und das gelingt uns nicht gut zur Zeit.

Moderatorin [00:33:28]

Noch eine Frage zu den KI-Kapazitäten, inwiefern das vielleicht Marketing ist oder was sie wirklich können. Gerne an Herr Geiping. Gerade wir als mehr oder weniger Laien und die Journalist:Innen können das gar nicht immer einschätzen, ob jetzt eine Gefahr gerechtfertigt angekündigt ist oder ob es eben nur Marketing ist. Wie kann man das erkennen? Und wie schätzen sie das bei aktuellen Entwicklungen ein?

Jonas Geiping [00:33:56]

Ja, das ist eine schwierige Frage. Also ich glaube eigentlich kann man das nur im Nachhinein sagen, nachdem das Modell von der Öffentlichkeit oder von vielen Forschern und Nutzern getestet wurde. Und dann entscheiden, was ist Marketing, was ist wirklich. Aber bei den Modellen, die wir jetzt hier gesehen haben in den letzten zwei Monaten wie 5.5 oder Mythos, sehen wir auch den Effekt. Also wir sehen zum Beispiel die Reports von FFmpeg, von Mozilla Firefox, von cURL, wir sehen wirklich diese Maintainer von Open-Source-Software und den Report von Herrn Holz. Wir sehen wirklich, dass diese Schwachstellen gefunden werden, wo sie vorher nicht gefunden wurden. Und das ist eben auch ein Signal von der Außenwelt, dass viele von diesen Maintainern wirklich auch sagen: Ja, Mythos ist bei uns gelaufen und es hat wirklich eine Schwachstelle gefunden. Und diese Schwachstellen sind schon wirklich real, die existieren. Zwar nicht immer so viele, wie dann oft gezählt wird, weil die Firmen auch gerne überzählen und alles zählen, was irgendwie Schwachstelle sein könnte. Aber wir sehen schon wirklich einen Effekt auch in der realen Welt, in echter Software, so abseits von der Ankündigung von den Firmen. Was natürlich auch spannend ist, das passt zusammen mit dem Trend, den wir auch messen: Es gibt Messungen davon, wie lange Modelle laufen können. Aufgaben, für die Menschen vielleicht 30 Minuten Zeit benötigen bis hin zu Aufgaben für die Menschen vielleicht vier, acht bis zwölf Stunden benötigen. Also diese Zeit, die Modelle unabhängig arbeiten können, hat sich jetzt in den letzten zwölf Monaten exponentiell erhöht. Und das passt auch zusammen damit, dass die Modelle hier viel unabhängiger agieren können und selber nach Schwachstellen suchen können. Ohne, dass jetzt zum Beispiel wie vor einem Jahr noch, viel mehr Interaktion mit einem Forscher passieren muss oder mit jemandem, der wirklich sucht und sagt schau hier und wir machen das und dann das. Die Modelle sind jetzt mittlerweile viel autonomer und das passt damit zusammen, dass sie auch viel besser Schwachstellen finden können.

Moderatorin [00:36:10]

Aber sie führen noch nicht autonom Cyberangriffe durch, richtig?

Jonas Geiping [00:36:16]

Da kann vielleicht Herr Holz mehr dazu sagen, wie autonom diese Angriffe sind, die getestet wurden.

Thorsten Holz [00:36:22]

Also in unserem Benchmark war es komplett autonom. Also da haben die Modelle als Eingabe die Informationen bekommen, dass für einen gewissen Input das Programm crasht und dann haben sie einfach nur das Programm bekommen. Und die Aufgabe war, automatisiert ein Exploit dafür zu entwickeln, also wirklich eine Software, die dann diese Schwachstelle ausnutzt und dann operativ wirklich einen eigenen Code ausführen kann. Und da war dann überhaupt keine Interaktion und die Modelle hatten dann bis zu sechs Stunden Zeit, diese Aufgabe zu lösen, ohne dass von uns irgendjemand eingegriffen hat.

Moderatorin [00:36:56]

Aber da waren Sicherheitsvorkehrungen ausgestellt, habe ich das richtig verstanden?

Thorsten Holz [00:37:00]

Also wir hatten verschiedene Szenarien. Einerseits ohne Schutzmechanismen, da können wir selektiv...

Moderatorin [00:37:06]

Ohne Schutzmechanismen des Modells oder des Systems, das angegriffen wurde?

Thorsten Holz [00:37:10]

Des Systems. Also das System war dann ohne Schutzmechanismen und dann haben wir Schritt für Schritt auch weitere Schutzmechanismen eingeschaltet, um zu sehen, wie schwierig wird das denn, wenn eben auch mehr Schutz verfügbar ist. Findet das Modell dann immer noch Wege, um die Schwachstelle auszunutzen und wirklich einen Exploit zu bauen? Und die haben dann mehr Zeit benötigt, also mehr Tokens und mehr Toolcalls, die im Endeffekt gemacht wurden. Aber sie waren trotzdem immer noch in der Lage, zwar eine niedrigere Zahl an Schwachstellen komplett automatisiert auszunutzen, aber es war immer noch nicht null.

Moderatorin [00:37:44]

Noch eine Frage an Herr Rieck, und zwar über das Claude-Mythos-Preview-Modell. Was ist denn bekannt über die Schwachstellen, die da gefunden wurden? Sind die relevant? Wie viele sind das? Und erlauben die Aufdeckungen von Mythos auch generelle Rückschlüsse darauf, wie man Software zukünftig sicherer gestalten kann?

Konrad Rieck [00:38:05]

Also nach meinem Stand weiß man noch nicht so viel. Also es gibt diesen Blogpost von vor zwei Wochen. Es gibt verschiedene Firmen, die die Software jetzt testen. Es gibt die Studie von dem Kollegen Holz. Also sozusagen nach und nach setzen sich da so Bausteine zusammen. Die Frage, welche Schwachstellen da gefunden werden, ist aus meiner Sicht noch nicht beantwortet. Ich würde sagen, das ist eine der spannendsten Fragen. Man kann sagen, dass diese Modelle niemals alle möglichen Schwachstellen finden können. Also das kann ich zum Beispiel beweisen. Aber es könnten trotzdem fast alle, die wir so in der Praxis sehen, die uns begegnen, gefunden werden,

vielleicht aber auch nicht. Das heißt also, die Frage, wo sind die Grenzen der Technologie, ist, glaube ich, eine sehr spannende Frage, die wir in der Forschung beantworten müssen. Was war der zweite Teil der Frage?

Moderatorin [00:38:55]

Ob da generelle Rückschlüsse draus gezogen werden können, wie Software sicherer gemacht werden kann.

Konrad Rieck [00:39:02]

Also erstmal: Jede Schwachstelle, die man findet, erzählt einem etwas darüber, was schief gelaufen ist in der Entwicklung. Teilweise entstehen die Fehler auch eben aus der Komplexität von Software. Ich wollte mich vorhin nicht melden, aber wir haben auch über Geld geredet und dass das so teuer ist. Und das ist vor der KI-Zeit auch schon so gewesen. Also, wenn Sie eine sichere Software entwickeln wollen, ist das immer teurer, als wenn sie einfach mal schnell irgendwas entwickeln. Und das ist einfach so. Wir hoffen, dass wir viel lernen. Gleichzeitig muss aber die Softwareentwicklung auch mehr auf Sicherheit achten, wenn wir Sicherheit als wichtiges Qualitätskriterium haben wollen.

Moderatorin [00:39:49]

Und es ist auf jeden Fall teuer, da rein zu investieren. Herr Holz, inwiefern sollten diese höheren Ausgaben für Cybersicherheit von privaten Unternehmen bezahlt werden? Oder muss der Staat da aus Ihrer Sicht was tun?

Thorsten Holz [00:40:01]

Ich denke, der Staat ist wahrscheinlich da der falsche Ansprechpartner. Ich hatte eben das Beispiel von Mozilla, also einem Open-Source-Projekt oder eben Microsoft oder andere Firmen nutzen das gerade oder investieren auch viel da rein. Ich denke, der Staat kann jetzt in dem Kontext nicht viel machen. Er kann höchstens nur fordern, dass die Hersteller dafür sorgen, dass die Software sicherer wird. Also da haben wir diverse Initiativen, gerade auf EU- oder auf deutscher Ebene, dass einfach auch die Unternehmen mehr in die Pflicht genommen werden, mehr für Sicherheit zu tun. Auch Punkte wie eine Meldepflicht von Sicherheitsvorfällen und andere Aspekte spielen da rein. Also ich denke, der Staat kann regulierend eingreifen. Allerdings die Kosten, glaube ich, sollte dann schon die Firmen tragen.

Moderatorin [00:40:50]

Noch eine Frage an Sie, Herr Geiping. Und zwar ist es jetzt auch so, dass immer mehr Agentensysteme eingesetzt werden, dass sie zum Beispiel automatisch Emails beantworten und so weiter. Bietet das noch mal eine neue Art Sicherheitslücke oder ein Einfallstor für Cyberangriffe und Hacks? Diese vernetzten Systeme.

Jonas Geiping [00:41:13]

Genau. Also bis jetzt haben wir nur über wirkliche Software Vulnerabilities gesprochen. Also Probleme mit klassischer Cybersecurity. Aber natürlich haben wir auch zum Teil, je nachdem, wie Agenten eingesetzt werden, wie Modelle eingesetzt werden, Probleme, die wirklich Machine

Learning basiert sind. Zum Beispiel dadurch, dass Systeme irgendwie überzeugt oder ausgetrickst werden können. Es gibt viel Forschung zum Thema Prompt Injection, also zum Thema Angriffe, in denen dem Modell irgendwo in einem Text auf einer Webseite eine neue Instruktion gegeben wird. Und je nachdem, wie das gemacht wird, kann sogar das Modell irritiert werden oder kann dazu gebracht werden, einer neuen Instruktion zu folgen. Das klassische Beispiel ist, dass jemand herausfindet, wie man eine Email schreiben kann an diesen Agenten, sodass der Agent dann sagt, das mache ich, ich schicke deine Emails an alle Leute und dann lösche ich alle Beweise dafür. Das haben wir tatsächlich in der Praxis noch nicht so viel gesehen. Tatsächlich ist es gar nicht so einfach, einen Angriff gegen ein Maschine-Learning-Modell zu machen, der verlässlich genug funktioniert, dass jemand davon profitieren kann. Aber gerade, wenn wir uns überlegen, dass Modelle in Bereichen wie Customer Service oder in Bereichen, in denen die Modelle wirklich Entscheidungen autonom treffen dürfen, eingesetzt werden. Dann gibt es da auch Anreize für die Angreifer, nach Machine-Learning-Problemen zu suchen. Also wirklich nach Schwachstellen in der Logik des Modells selber intern und nicht Schwachstellen in der Software um das Modell. Das ist natürlich etwas, was wir weiter verfolgen und wo immer ein bisschen auch die Frage ist, wie sich weiterentwickelt wird, ob die Modelle sicherer werden in dieser Hinsicht. Ob die Modelle weiterhin angreifbar bleiben, das werden wir sehen.

Moderatorin [00:42:58]

Wird sich mit der Zeit noch zeigen. Wir haben es gerade schon angesprochen. Einmal, dass es da auch Fehler in Software geben kann. Da haben wir noch eine passende Frage zu bekommen, und zwar an Sie vielleicht, Herr Rieck, ob es vielleicht ein Anreiz wäre, Softwareentwickler:Innen für Schäden durch Sicherheitslücken haftbar zu machen. Dass sie dann vielleicht bessere Software programmieren?

Konrad Rieck [00:43:22]

Also das kann ich nicht beantworten, die Frage. Aber ich habe kurz bei der Frage, ob der Staat da Geld geben sollte, an andere Dinge gedacht. Also zum Beispiel, wenn ich ein Busunternehmen gründe und mein Bus ist unsicher und der macht andauernd Unfälle, dann fände ich das verrückt, wenn der Staat sich darum kümmern würde, dass die Busse da sicherer werden. Das muss schon das Unternehmen machen. Das gehört zu dem Geschäftsmodell und das sind Kosten, die Teil des Geschäftsmodells sind. Die kann man sich zwar wegdenken, aber in der Realität sind sie da. Wer jetzt haftbar ist, das kann ich nicht sagen. Das hängt immer auch davon ab, wie die Dinge mit den Geschäftsbedingungen geregelt sind und so weiter. Es ist alles unheimlich kompliziert. Das wollte ich vorhin auch schon sagen, als es um die Note für Deutschland ging: Es gibt Software, die in Deutschland entwickelt wird, aber die Art und Weise, wie Software generell entwickelt wird, ist überhaupt nicht an Ländergrenzen gebunden. Also von daher kann man Noten nicht für Länder vergeben, sondern vielleicht einfach für Arten von Firmen oder Arten von Software. Es ist ein internationales Geschäft und wie die rechtliche Verantwortung ist, das entzieht sich meiner Kenntnis.

Moderatorin [00:44:34]

Aber vielleicht nicht Herr Geipings Kenntnis. Was möchten Sie dazu sagen?

Jonas Geiping [00:44:37]

Nein, also das entzieht sich auch meiner Kenntnis. Aber vielleicht ein Nachtrag zu dem Thema, was Deutschland oder Europa vielleicht machen sollen. Was natürlich hier schade wäre, wäre, wenn wir

durch diese Problematik viel mehr Zentralisierung von Software sehen würden. Also wenn es jetzt für kleine Unternehmen in Deutschland nicht mehr möglich ist, sichere Software zu erzeugen und das alles eingekauft werden muss, zum Beispiel aus den USA. Das wäre natürlich für Deutschland irgendwie auch ein Verlust. Vielleicht kann man da was machen oder darüber nachdenken, wie auch diese eigene Softwareindustrie geschützt der auch verbessert werden kann, um nicht den Effekt zu erzeugen, dass man überhaupt keine Software mehr hier bauen kann, die sicher sein kann. Es wäre irgendwie schade, wenn das alles nur noch Microsoft ist oder zum Beispiel AWS, eben weil das Unternehmen sind, die viel mehr Geld investieren können, um ihre Software abzusichern. Das wäre ein Verlust.

Moderatorin [00:45:34]

Das ist sicherlich auch nicht gut für die digitale Souveränität. Haben Sie denn da zufällig eine Idee, wie man dafür sorgen kann, dass wir eben nicht so abhängig sind von großen Unternehmen? Alle von Ihnen gerne, wenn jemand eine Idee hat. Sieht tatsächlich nicht so aus. Oder Herr Holz?

Thorsten Holz [00:45:52]

Das ist ein Punkt, den wir bis jetzt noch nicht angesprochen haben. Aber wir sind jetzt aktuell sehr abhängig, vor allem von US-Anbietern und die starken offenen Modelle kommen auch alle aus China. Europa hat mit Mistral noch eine Firma, die allerdings auch nicht komplett konkurrenzfähig ist. Die sind auch schon deutlich hinterher. In Deutschland war AlephAlpha eine Zeit lang gehypt, die haben in dem Kontext keine große Bedeutung mehr. Deshalb, was Europa da auch leider wieder mal verschlafen hat, ist, da den Anschluss zu finden. Ich denke, da müssen wir schauen, wie man das konzentriert tun kann. Jetzt in den kommenden Jahren dann noch mal etwas aufzuholen, anstatt zu sagen, wir sind komplett raus. Wobei man auch sagen muss, dass in den USA gerade sehr große Summen im zwei-, dreistelligen Milliardenbereich investiert werden, um Rechenkapazitäten aufzubauen. China baut da auch sehr viel gerade auf. Und da fehlt in Europa einerseits der Wille und andererseits auch die operative Umsetzung. Also ich glaube, da müssen wir uns kritisch damit auseinandersetzen: Was für technologische Abhängigkeiten haben wir zu den USA oder auch zu China? Wie können wir dafür sorgen, dass wir in Europa auch mehr Kompetenz aufbauen, um einfach in Zukunft nicht komplett abgehängt zu werden? In dem Kontext habe ich gerade einen Link reingepostet. Wir hatten mit verschiedenen Forschenden aus Europa einen offenen Brief veröffentlicht, in dem wir auch fordern, dass Europa da koordiniert mehr Kompetenzen aufbauen muss. Gerade in dem Schnittbereich zwischen AI und Cybersicherheit, um da nicht komplett technologisch abhängig zu sein von anderen Ländern.

Moderatorin [00:47:39]

Den Link wird ein Kollege von mir sammeln und zur Verfügung stellen. Den müssten Sie aber auch schon haben aus der Einladungsmail beziehungsweise der Remindermail, liebe Journalistinnen und Journalisten, für dieses Briefing. Sie können da auf jeden Fall drauf zugreifen. Ich habe für jeden von Ihnen jetzt noch eine mehr oder weniger abschließende Frage vorbereitet und würde dann gerne mit Ihnen, Herr Rieck, anfangen. Was ist Ihre Prognose, wie sich die Cybersicherheit in Zukunft durch KI noch weiter verändern wird? Wir haben jetzt schon gesagt, Angriffe und Verteidigung sind beides sehr viel schneller möglich, aber was ist Ihre Prognose für die Zukunft, wie es da weitergeht?

Konrad Rieck [00:48:15]

Es gibt eine große Unbekannte. Das hängt davon ab, wie viel Software in der Zukunft durch KI geschrieben wird. Denn dann kommen wir zu dem Punkt, an dem KI Schwachstellen in der Software findet, die KI schreibt. Und das ist mir völlig unklar. Was ich aber noch gerade zu dem Punkt loswerden wollte, und ich weiß nicht, ob das dazugehört, aber ich glaube, damit Europa mithalten kann, braucht es sehr, sehr viel Geld. Und ich habe den Eindruck, dass das sowohl den Regierungen als auch den Firmen in Europa nicht klar ist, wie viel Geld in Amerika investiert wurde und dass man natürlich einen Nutzen von solchen Milliardeninvestitionen hat und dass wir zum Beispiel nicht nur durch Manpower oder schlaue Ideen da mithalten können. Und solange das Geld nicht investiert wird, sehe ich da ein bisschen schwarz.

Moderatorin [00:49:04]

Haben Sie da Pi mal Daumen Größenordnungen, wie viel in den USA investiert wird, wie viel in Europa investiert wird und wie viel investiert werden soll?

Konrad Rieck [00:49:10]

Das weiß ich leider nicht. Aber wenn ich an AlephAlpha denke, die hatten, glaube ich, so was wie 500 Millionen. Und das ist vielleicht ein Zehntel oder ein Zwanzigstel von einer amerikanischen Firma. Und wie soll das dann funktionieren? Die Firma gibt es schon noch, aber die haben ihre Pläne, selber solche Foundation-Modelle from Scratch zu bauen, meines Wissens auch eingestellt, sodass wir da auch nur Nutzer von Technologie sind, die andere entwickeln.

Moderatorin [00:49:47]

Dann gerne die gleiche Frage auch noch an Herr Holz. Wie ist Ihre Prognose für die Zukunft? Und wann werden wir große Veränderungen durch KI als Endnutzer spüren?

Thorsten Holz [00:49:58]

Das ist natürlich ein Blick in die Glaskugel, weil vor sechs oder zwölf Monaten hätte ich es noch ziemlich für undenkbar gehalten, dass die Modelle in der Lage sind, eine Kernel-Schwachstelle oder eine Schwachstelle in einem Browser automatisiert auszunutzen. Jetzt haben wir allerdings schon Modelle, die das können. Und es ist unklar, was jetzt in sechs bis zwölf Monaten möglich ist. Und da stimme ich auch mit Herrn Rieck komplett überein. Es wird sowohl auf der Angriffs- als auch auf der Defensivseite sehr viele Veränderungen geben. Wir haben jetzt auf einmal wieder sehr viele neue Schwachstellen, die wir finden, die schon vorhanden sind, die wir aber noch nicht kennen, mit denen wir uns auseinandersetzen müssen. Und das ist eben die Frage: Wie wir das Ganze sich in Zukunft noch verändern? Und ich gehe davon aus, dass wir jetzt erstmal innerhalb der nächsten eins, zwei, vielleicht drei Jahre so eine etwas schwierige Phase haben, weil es einfach sehr viele neue Schwachstellen gibt, für die wir erstmal lernen müssen, wie können wir damit umgehen, wie können wir die patchen, wie können wir dafür sorgen, dass auch operativ die Patches eingespielt werden und unsere Systeme sicherer werden? Und was danach passiert, ist dann so ein bisschen unklar, wie viel stärker oder viel besser werden die Modelle in der Zukunft, wie kann man die auch für defensive Mechanismen besser einsetzen? Und haben wir irgendwann auch noch mal einen größeren Durchbruch? Meine typische Analogie ist, als damals AlphaGo rauskam, also ein maschinelles Lernmodell in der Lage war, den weltbesten Go-Spieler zu schlagen, da gab es in einer dieser Partien diesen Move 37, den 37. Zug. Das haben Menschen erstmal nicht verstanden, allerdings dann später schon, als sich herausgestellt hat, dass das eigentlich der

spielentscheidende Zug war. Und hier die Frage, was ist das ganze im Bereich Cybersecurity? Wann sind Modelle in der Lage, komplett neue Angriffstechniken zu entwickeln oder eben auch komplett neue Schutzmechanismen zu entwickeln, die dann Angriffe vielleicht unmöglich oder auf jeden Fall deutlich schwieriger machen. Und genau diese Art von Durchbrüchen, das ist eben spannend und nicht wirklich absehbar, ob sowas jetzt innerhalb der nächsten ein, zwei oder drei Jahre passiert im Bereich Cybersecurity.

Moderatorin [00:52:07]

Herr Geiping, können Sie vielleicht abschätzen, wann Modelle so gut werden? Und kann man diese Modelle dann überhaupt noch sicher machen, wenn Sie diese ganzen Fähigkeiten besitzen?

Jonas Geiping [00:52:20]

Ich glaube, das ist noch eine größere Glaskugel. Wie sich diese Modelle in den nächsten, sagen wir mal, zwölf Monaten entwickeln werden. Das ist wirklich unklar. Wir schauen halt, was passiert. Das einzige, was wir sehen, ist, dass die Verbesserung der Modelle schon irgendwie proportional ist zur Investition. Mythos ist trainiert auf der neuesten Generation von Nvidia-Karten. Auf den größten Clustern, die die Firmen gebaut haben. Und was wir sehen können, ist eine Proportionalität, ist ein Faktor zehn oder hundert in Rechenleistung und Geld, je nachdem. Das ist der Faktor, der sozusagen nötig war, um von der letzten Generation von Modellen auf diese Generation aufzusteigen. Aber viel mehr Faktoren an Compute sind physikalisch ab einem gewissen Punkt gar nicht so einfach. Vielleicht wird es dann schwieriger, die Modelle weiterhin so schnell zu verbessern. Jetzt sind wir bei Gigawatt Clustern, die irgendwo stehen. Es kostet zum Beispiel 50 Billionen, dieses Cluster zu bauen, um das Modell zu trainieren. Das können wir auch nur so oft wiederholen, dieses Mal zehn hier. Andererseits könnte es auch die große Frage sein, ob die Modelle auf andere Arten besser werden, die wir schwer vorhersagen können. Es ist gar nicht so einfach. Und dann die Frage, wie sich die Sicherheit der Modelle selber entwickelt. Es kann natürlich sein, dass die Modelle noch besser werden, wir sie besser einsetzen und sie mehr Kontext haben oder bessere Erinnerungssysteme. Oder, dass die Modelle auch selber verstehen, was der Benutzer von ihnen will und dass Angriffe für Durchschnittsbenutzer schwieriger werden, weil das Modell eher versteht, was passiert und eher diesen Angriff abwehren kann oder sagen kann, das machen wir nicht, das ist Quatsch. Aber da müssen wir schauen, wie es sich entwickelt. Das ist wirklich schwer vorherzusagen.

Moderatorin [00:54:17]

Okay. Dann enden wir, glaube ich, mit einem sehr offenen Ende. Und wir müssen sehen, was die Zukunft bringt. Wie das so oft ist. Ich bedanke mich bei Ihnen, liebe Experten, und auch bei Ihnen, liebe Journalistinnen und Journalisten, für die Fragen, die Sie gestellt haben, für Ihre Aufmerksamkeit. Eine Audio-, eine Videodatei und ein maschinell erstelltes Transkript stellen wir so schnell wie möglich jetzt nach Ende dieses Meetings zur Verfügung. Sie finden die Dateien über die Links in der Einladungsmail, die Sie Anfang der Woche bekommen haben, und der Remindermail von heute Morgen. Ein redigiertes Transkript stellen wir über dieselben Links dann auch im Laufe des Tages zur Verfügung, sobald wir es haben. Bis zum nächsten Mal. 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