# KI Trends reflektiert: Technologische Entwicklungen zwischen Nobelpreis und Regulierungen

Ausgesprochen digital. Der Podcast für digitale Trends.

## Intro

[00:00:06.770] - Steffen Wenzel

Herzlich willkommen bei Ausgesprochen digital. Wenn es auch schon ein wenig spät ist, hoffen wir, dass Sie gut ins neue Jahr gestartet sind. Und wir freuen uns natürlich auch sehr, dass Sie wieder bei uns zuhören. Frank Schönefeld ist heute zu Gast, insgesamt zum sechsten Mal. Und jetzt zum Dritten Mal treffen wir uns immer so plus minus rund um den Jahreswechsel herum, um über die neuesten Entwicklungen zum Thema künstliche Intelligenz zu reden. Wir nehmen diese Folge am 9. Januar auf. Die MMS feiert heute ihr 30 jähriges Jubiläum. Frank, bist du eigentlich von Anfang an dabei?

[00:00:44.240] - Prof. Dr. Frank Schönefeld

Also ebenfalls gesundes neues Jahr von mir in die Runde der Zuhörenden. Ich habe 27 Jahre auf dem Buckel mit der MMS, also nicht ganz von Anfang an. Ich bin gekommen 1998, da war die MMS drei Jahre alt. Ironischer Weise, könnte man sagen mit Bezug auf unser heutiges Gespräch, bin ich von einer Firma gekommen, die neuronale Netze für Zeitreihenvorhersage und für Gesichtserkennung verwendet hat und wie das Schicksal manchmal so spielt, 20 Jahre später erlebt diese Technologie eine Wiederauferstehung, Wiederbelebung. Und mit Deep Learning revolutioniert sie eben die gesamte künstliche Intelligenz. Da habe ich vor 25 Jahren mal angefangen oder vor 30 Jahren nunmehr.

[00:01:31.100] - Steffen Wenzel

Ja und da ist anscheinend ja auch ein bisschen was passiert in den letzten 25 Jahren. Lass uns noch mal kurz auf das Jahr 2024 zurückblicken. Du hast im letzten Podcast den Launch von Chat GPT mit der Kambrischen Explosion verglichen, also diesem erstaunlichen Ereignis vor über 500 Millionen Jahren, als plötzlich die Artenvielfalt und die Komplexität des Lebens auf der Erde enorm anstieg. Bleibst du bei diesem Vergleich oder hat sich da etwas in deiner Betrachtung der Dinge verändert.

[00:01:59.840] - Prof. Dr. Frank Schönefeld

Die kambrische Explosion steht ja für ein Ereignis, was in relativ schneller Geschwindigkeit abläuft und was dann aber nachhaltig und dauerhaft eine Veränderung herbeiführt. Und ich denke, diese Attribute treffen für das, was wir gerade erlebt haben und noch erleben, absolut zu. Wir datieren die kambrische Explosion der KI mit dem November 2022 mit dem Launch von GPT 3.0, und seitdem sind jetzt zwei Jahre vergangen und ich denke, es war schnell und es hat eine dauerhafte, nachhaltige Veränderung, nachhaltig im Sinne von anhaltend eine anhaltende Veränderung geführt. Also ich glaube, das Bild trägt nach wie vor. Ich könnte eine Ergänzung machen im Sinne ich glaube, es hat noch keine Nacheruptionen gegeben oder Nachexplosion, sondern wir sind immer noch Zeuge dieser ersten Welle, die um den Planeten schwappt und gerade die letzten Inseln erreicht. Aber wir basieren immer noch auf den damals verwendeten Technologien, Transformer Architekturen, Attention Mechanismen usw. Und ich glaube, wir brauchen aber auch den nächsten Schritt, eine nächste Eruption, Explosion, um die damit auch in Zusammenhang stehenden Herausforderungen zu meistern. Ich reflektiere insbesondere auf den hohen Energieverbrauch der aktuellen KI Modelle.

## Nobelpreise für Informatiker

[00:03:33.140] - Steffen Wenzel

Da werden wir auch noch mal zukommen. In dem Bild noch mal zu bleiben, ich glaube, es ging so um 10 Millionen Jahre bei der kambrischen Explosion, wo sich alles entwickelt haben. Jetzt haben wir gerade mal so 2 bis 3 Jahre hinter uns. Lass uns deswegen noch mal schauen, was ist in den letzten Jahren passiert? Was war für dich so das Bedeutende, gerade im Jahr 2024? Welche Neuentwicklung gibt es auch, die dich besonders beeindruckt haben?

[00:03:58.300] - Prof. Dr. Frank Schönefeld

Ich fange mal mit einem Ereignis an, was vielleicht nicht jedermann als Nummer eins auf dem Zettel hat, aber ich würde doch die Nobelpreise erwähnen. Zum ersten Mal in der Geschichte der Nobelpreise. Und die haben auch ihre 120 Jahre auf dem Buckel, sind zwei Informatiker ausgezeichnet worden, zwei Informatiker in einem Jahr. Und natürlich gibt es keinen Nobelpreis für Informatik. Man muss diese Preise in die alten Kategorien pressen. Aber der Nobelpreis für Physik ist an, unter anderem, an Geoffrey Hinton verliehen worden, einen der Pioniere der neuronalen Netze. Und der Nobelpreis für Chemie ist ein Demis Hassabis verliehen worden für seine bahnbrechenden Arbeiten um Struktur vorhersagen, wie sich Eiweiße final falten werden. Zu ehren und zu würdigen. Basierend auf dem Modell AlphaFold 2, was wiederum eine Verwandtschaft zu AlphaGo Zero hat. Und insofern kann man sagen, dass die Fähigkeit, gut zu spielen, Spieltheorie sich jetzt hin bis zum Nobelpreis ausgezeichnet hat. Also eine der bemerkenswertesten Ereignisse des Jahres 24 für mich. Erstens Informatik, wenn auch in die Kategorien Physik und Chemie verpackt. Können wir darüber hinwegsehen.

[00:05:26.100] - Prof. Dr. Frank Schönefeld

Ich möchte aber auch die Schnelligkeit des Nobelpreiskomitees eigentlich loben und hervorheben. Die Erkenntnisse von Demis Hassabis sind in etwa drei Jahre alt gewesen und so eine Sache so schnell mit einem Nobelpreis zu würdigen, halte ich für außerordentlich. Also ich glaube, die gesamte Computer Science, die gesamte Deep Learning Community können sich geehrt fühlen und es ist eine tolle, tolle Errungenschaft.

[00:05:54.720] - Steffen Wenzel

Wir werden ja sicherlich in diesem Podcast oder in der Zeit, die wir heute haben, noch mal drüber reden. Aber es wurde natürlich auch und das war das Besondere an diesem Nobelpreis oder der Auszeichnung auch vor KI gewarnt durch einen der Preisträger. Wie schätzt du diese Tatsache ein?

[00:06:13.920] - Prof. Dr. Frank Schönefeld

Auch unter dem Aspekt gibt es, glaube ich, kaum einen Würdigeren als Geoffrey Hinton. Er steht wie kein Zweiter, erstens für wirklich technologischen Fortschritt, aber gleichzeitig eben die Erkenntnisse: Was kann daraus alles folgen? Und auch die, er hat sich die Fähigkeit offengehalten oder wiedereröffnet, auch kritisch den eigenen Ergebnissen gegenüber zu stehen. Er war ja Chief Scientist bei Google, hat das Unternehmen verlassen, um auch keine Corporate Zwänge auferlegt zu bekommen, um zu sagen: KI ist gut, up to the point, bis zu einem gewissen Punkt, aber wir müssen auch Maßnahmen, Regulierungen treffen, um die schlimmsten Auswüchse zu beschränken.

## Rückblick und Entwicklungen

[00:07:01.770] - Steffen Wenzel

Gut, dann lass uns doch mal zusammenfassen. Was sind so die wichtigsten Entwicklungen deiner Meinung nach? Du hast Google jetzt gerade schon genannt, einen der Player, die Big Five in den USA. Aber es gibt ja auch noch andere in Europa. In Deutschland ist sogar etwas dazugekommen. Kannst du uns das noch mal zusammenfassen?

[00:07:19.110] - Prof. Dr. Frank Schönefeld

Ich versuche es Mal zusammenzufassen. Und das ist eine Herausforderung an sich, weil die Explosion, um das Bild noch mal zu bemühen, wirklich unendlich viele Wellen aufgeworfen hat. Manche kräuseln sich so leise im Wind und manche haben richtig Tsunamicharakter angenommen. Von den Magnificent Seven, wie man sie nennt, die glorreichen Sieben in den USA sind ja mindestens sechs direkt mit dieser Entwicklung verbunden. Aber auch Tesla hat natürlich massives Interesse, sich dort weiterzuentwickeln. Spitzenreiter in dem Umfeld bleibt Nvidia, die die Hardwarerevolution in diesem Gebiet anführen. Gerade auf der CES in Las Vegas in dieser Woche erst annonciert, ein ein Desktopcomputer für etwa 3.000 $, der eine massive Beschleunigung sämtlicher KI-Prozesse anbieten wird. Nvidia hat sich allein im letzten Jahr noch mal um 200% im Börsenwert gewachsen und führt damit die wirtschaftliche Performance an. Aber auch Google, tolle Resultate, 50% Wachstum und die anderen folgen dann in gewissem Abstand. Worauf basiert das? Bei Nvidia ist es klar die die Hardwareüberlegenheit. Den Desktop Computer Project Digital hatte ich schon angesprochen, aber insbesondere im Servergeschäft, wo die großen Netze trainiert werden. Die H1 Architekturen und jetzt neu die die Blackwell Architekturen.

[00:08:58.760] - Prof. Dr. Frank Schönefeld

Damit führt Nvidia unangefochten dieses Segment an. Die anderen, die battlen sich ein bissel stärker, aber sie haben die Claims so abgesteckt, dass sie sich nicht zu sehr ins Gehege kommen. So könnte man sagen: Jeder von ihnen hat mindestens ein großes Sprachmodell, meistens einen ganzen Zoo großer Sprachmodelle, die spezialisiert sind. Und in der Leistungsfähigkeit, es ist sehr interessant, das zu beobachten, nehmen Sie sich nur wenige Zehntel Prozentpunkte oder Prozentpunkte und das wechselt immer. Wir können sagen, dass Gemini 2.0, Gemini Flash ungefähr ex aequo mit GPT -o1, Omega und o1 praktisch liegt. Wer ein bisschen aus einer europäischen Brille noch durchs Raster fällt, ist Anthropic, der Partner von Amazon. Aber wenn man ganz genau die Benchmarks liest, sieht man, dass Anthropic eigentlich führt. Anthropic für die meisten der Benchmarks an, zwar auch nur um Zehntel Prozente oder weniger Prozente, aber insofern ist das leistungsfähigste Modell ist Claude sonnet, Claude haiku, Claude opus, welche Variante auch immer man nehmen mag. Und das spiegelt sich in der europäischen Wahrnehmung noch nicht so richtig wieder.

[00:10:24.280] - Prof. Dr. Frank Schönefeld

Wir werden sehen, wohin das führt. Gestern las ich die Nachricht, dass der vorbörsliche Wert, den man Anthropic zumisst, bereits höher ist als der von Mercedes Benz in Deutschland. Ich denke, zugespitzter kann man das Dilemma der deutschen Wirtschaft auch nicht auf den Punkt bringen. Wer überraschenderweise eine ganz tolle Rolle spielt, ist Meta mit seinen scheinbar Open Source Modellen. Also scheinbar sage ich deswegen die Lizenz Bedingungen sind schon zur Nutzung geeignet, aber es ist natürlich nicht Quelloffen. Und das sind natürlich die Kritikpunkte, die man noch anbringen könnte. Aber die LLaMA Modelle, letzte Modellvariante LLaMA 3.3 und auch die große Variante mit inzwischen 405 Milliarden Parametern trainiert, spielen eine wesentliche Rolle. Man kann sagen, dass pro Tag und das sind jetzt die letzten zwei Jahre 1 Million Versionen von LLaMA pro Tag heruntergeladen wurden. Und da merkt man, welche Verteilung, welche Distribution da natürlich dahinter steckt. Also Meta hat auch einen guten Weg gefunden, sich in diesem Geschäft zu positionieren. Der First Mover Advantage ist ganz klar bei GPT, OpenAI und Co. Man kann zum Beispiel die mobile Nutzung von GPT mit der mobilen Nutzung von Gemini vergleichen und da wird man auf einen Faktor fünf, Faktor sechs zugunsten von GPT kommen, obwohl die Leistungsfähigkeit eigentlich ex aequo, ex aequo liegt.

## Edge Modelle und Kernkraft als Lösungen für den Energiehunger?

[00:12:04.250] - Prof. Dr. Frank Schönefeld

Also wahnsinnige Fortschritte, Verbreitung, Modelle in allen Abstufungen. Über Edge Modelle haben wir noch nicht gesprochen, aber können wir im Zuge der Energiediskussion. Da ist ein Edge Modell natürlich eine mögliche Antwort darauf. Können wir darauf noch kommen. Wenn wir einmal dabei sind. Ein Edge Modell wäre natürlich eins, was nicht zwingend auf Server, Connectivity, auf Connectivity in die Cloud hinein angewiesen ist, sondern vor Ort in einer lokalen Installation laufen kann und nur mit den

begrenzten Ressourcen dennoch zu solchen hohen Intelligenzleistungen praktisch fähig ist. Also Edge für außerhalb der Connectivity und Cloud Verbindung. Und naturgemäß kann man dann nicht so viel Power dort reinstecken. Also so viel Macht, so viele Parameter. Das sind typischerweise Modelle mit weniger Parametern, vielleicht 1 Milliarde Parametern, gegebenenfalls noch weniger Parametern und eher zugeschnitten auf einen speziellen Zweck. Und das hilft natürlich auch, Energie zu sparen. Es kann lokal ausgeführt werden, kann damit aber auch Schnelligkeitsvorteile entwickeln, weil die gesamte Latenz und Server Latenz ebenfalls wegfällt.

[00:13:21.700] - Steffen Wenzel

Aber die Energieersparnis erfolgt quasi im Betrieb, aber nicht in dem Sinne, wie wie das Lernmodell entsteht?

[00:13:30.220] - Prof. Dr. Frank Schönefeld

Das ist fast richtig. Sagen wir mal so. Natürlich ist das Training, ist abhängig von dem Gesamtsetup. Die Trainingsdauer bestimmt sich über die Menge an Trainingsdaten, die ich hineingebe. Und wenn ich ein zugeschnittenes Modell trainiere, gebe ich natürlich auch weniger Daten in das Modell hinein. Und sie ist natürlich auch abhängig: Wie breit ist die Architektur aufgebaut, welche context windows size, um mal einen Fachbegriff zu verwenden, sind erlaubt? Wie viele Schichten sind in meinem Encoder oder Decoder verbaut? Und das sind die Parameter generell. Wenn man jetzt aber einmal durch Zufall bei dem Energiethema gelandet sind, muss man sagen, es hat ein großer Wechsel von klassischer KI, analytischer KI, zu generativer KI stattgefunden. Während wir bei analytischer KI noch sagen konnten: Training teuer, Inferenz oder Anwendung preiswert, überschaubar, im Aufwand. Haben wir jetzt den Fall Generative KI: Training sehr teuer, Anwendung immer noch teuer. Und das macht den großen Unterschied aus. Man geht davon aus, dass sich der Energieverbrauch für KI in Rechenzentren bis zum Jahr 2030 verdoppeln, höchstwahrscheinlich sogar verdreifachen wird und dass dann etwa bis zu 6 % des Primärenergieverbrauchs allein in KI gesteckt wird. Das macht noch mal die Dimension deutlich.

[00:15:04.590] - Steffen Wenzel

Des weltweiten Energieverbrauchs?

[00:15:05.970] - Prof. Dr. Frank Schönefeld

Des weltweiten Energieverbrauchs. Und da sich die großen Anbieter natürlich trotzdem zu einer Klimaneutralität, CO2 Neutralität, verpflichtet haben, bleibt ihnen nur ein Ausweg in dem Umfeld, nämlich auf Energie Supply zu setzen, welches wenig CO2 oder kaum CO2 erzeugt. Und da kommt überraschenderweise die Kernkraft wieder ins Spiel. Und Google hat Verträge geschlossen, wo bis zu sechs oder acht neue Kernkraftwerke den Energiehunger der Google KI stillen sollen. Da geht es um sogenannte small power modules, also kleine Reaktoren, die auch nicht ganz die Gefahr ausstrahlen, im wahrsten Sinne des Wortes, wie wir sie von den anderen Kernkraftwerken kennen. Aber es bleibt natürlich trotzdem eine nukleare Stromerzeugung. Microsoft hat einen Vertrag geschlossen, in dessen Zuge Three Miles Island wieder in Betrieb genommen werden soll. Und jeder, der sich mit Kernkraftunfällen beschäftigt hat, dem wird Three Miles Island noch in den Ohren klingen. Das war der größte anzunehmende Unfall, der seinerzeit in den USA stattgefunden hat. Also diese KI hat Folgen schon jetzt und vielleicht auf Gebieten, die wir gar nicht so im Blick hatten.

[00:16:28.290] - Steffen Wenzel

Ja, ich finde das gut, dass du das ansprichst. Und ich möchte es auch noch mal ganz kurz vertiefen, weil es ist natürlich ja auch eine Haltungsfrage, die wir als Personen haben, wenn wir KI benutzen, dass wir uns die Frage stellen sollten. Ich habe mal jetzt gelesen, Tiktok hat mittlerweile, alleine Tiktok mit seinen Streamingdiensten, ist jetzt kein KI Thema, aber dennoch macht es das Beispiel ein bisschen deutlicher, verbraucht so viel Energie wie ganz Griechenland. Glaubst du, das wird auch noch mal entscheidend sein für die großen Player? Auch im Bewusstsein, dass wir uns als Bevölkerung ja immer weiter damit beschäftigen müssen, wenn wir den Klimawandel quasi vor der Haustür haben und auch immer mehr spüren. Glaubst du, das wird ein Entscheidungskriterium sein, welche Anbieter wir in Zukunft wählen?

Dass die damit einen Vorteil erzielen, wenn sie dann wirklich klimaneutral oder mit alternativen Energien ihre KI betreiben können?

[00:17:20.600] - Prof. Dr. Frank Schönefeld

Ich denke, am Ende wird es über die, die die Kosten entschieden und das wird erst mal eine Menge determinieren. Und dann muss man natürlich dieses small modules, die small atomic oder nuclear modules muss man sich erst mal anschauen. Noch existieren die ja nur auf dem Papier. Und welche Leistungsfähigkeit haben, welche Kostenstruktur und welche Risiken gehen damit einher? Das ist im Moment nur unvollständig abbildbar. Ich, ich glaube nicht, dass es über moralische Kategorien entschieden wird, sondern eher über rein Wirtschaftliche.

## Anwendungsfälle

[00:17:57.200] - Steffen Wenzel

It's the economy, stupid, wie immer. Das kennen wir jetzt schon einige Jahre. Lass uns mal ein bisschen in die Anwendungsfälle schauen. Weil es gibt ja jetzt, natürlich kennt jeder Chat GPT. Meine Tochter benutzt das jetzt sogar schon für Hausaufgaben. Ich übrigens auch, weil ich dann Mathe Geometrie nicht mehr drauf habe. Also es ist wirklich eine super Hilfe. Also es ist wirklich in der Gesellschaft, glaube ich angekommen. Aber es gibt natürlich eine Vielfalt von Anwendungsmöglichkeiten. Kannst du uns da mal ein bisschen noch mit auf die Reise nehmen?

[00:18:27.590] - Prof. Dr. Frank Schönefeld

Ja, ich denke, wir können durchaus zwei große Fälle unterscheiden. Das eine ist so wie du es schilderst, deine Tochter, im Bekanntenkreis etc. Das wäre also die, im E-Commerce haben wir früher B2C Nutzung gesagt, also eher die private Nutzung, also im privaten Bereich und dort nehmen wir wahr, wie es wirklich schleichend sozusagen hinein dringt. Jeder hat die die erste App oder es ist in Apps integriert, die man lange genutzt hat und damit nutzt man es schleichend. Ich halte ja Vorlesungen an der HTW Dresden und zu KI und jedes Jahr frage ich die Studenten: Wer hat in der letzten Woche generative KI genutzt? Und ich kann mich noch an die Vorlesung vor einem Jahr erinnern. Da waren das so ein paar vereinzelte Hände, die hoch gingen. Und im letzten Dezember, also vor einem Monat, habe ich das erneut gefragt und da waren es vereinzelte Hände, die nicht hoch gingen. Also es waren eher 85% der Studenten die, die das nutzen. Und natürlich nutzen auch wir das bei Ausarbeitungen. Eine ganz simple Sache habe ich im persönlichen Bereich erlebt.

[00:19:42.230] - Prof. Dr. Frank Schönefeld

Es musste in eines dieser englischsprachigen HR-Systeme eine Selbsteinschätzung eingegeben werden und auch noch in einer in Englisch, also nicht in der deutschen Muttersprache. Ja, was macht man in einem solchen Falle? Da lässt man sich, sich helfen. Ja. Das ist das und das sorgt natürlich dafür, wenn man mit diesen Erfahrungen aus dem Privatbereich kommt, dann überträgt man das Gelernte oder auch Gewollte natürlich auch ins Geschäft und fragt sich: Warum geht das da noch nicht? Und das ist so ähnlich der Effekt, wie wir dann auch im E-Commerce hatten. Wir waren alle Amazon geschult und haben uns gefragt: Warum zur Hölle können das die internen Einkaufssysteme nicht so intelligent abbilden? Und das hat natürlich sukzessive zu einem Druck auf diese Plattform geführt. Und inzwischen sehen sie fast alle aus wie Amazon sozusagen. Und das würde ich auch hier im Geschäftsbereich erwarten. Da liegen Zahlen dazu vor. Bei den Fortune 500, also den 500 größten Unternehmen der Welt, da haben wir eine Quote, die jenseits der 92% liegt, die ganz gezielt künstliche Intelligenz einsetzen. Es gibt eine wunderschöne Sammlung von Use Case Fällen von Google.

[00:21:02.600] - Prof. Dr. Frank Schönefeld

Kann ich wirklich empfehlen, über 180 Use Cases von KI wirklich querbeet vom Marketing angefangen bis hin zur Einkaufsunterstützung, bis hin zur Marketing Content Generierung. Also in Zukunft können, was wir heute machen können, wir generieren lassen. Also das ist sehr breit. Wenn man jetzt einen Schritt

zurückgeht und sagt Wie sieht es denn im Mittelstand aus? Da merkt man, dass es einen Zusammenhang gibt zwischen digital readiness. Also wie weit sind die Hausaufgaben in der digitalen Transformation gemacht? Und wer dort gut ist, der kann einfach nahtlos als letzte Etappe natürlich jetzt die KI Etappe anschließen. Und wer die digitalen Transformations-Hausaufgaben nicht gemacht hat, der wird natürlich noch mehr Mühe haben, die die aktuelle Etappe KI da dranzuhängen und zu gehen.

## Nachfrage und Tendenz im Mittelstand

[00:21:56.740] - Steffen Wenzel

Also das heißt es kommt jetzt auch im Mittelstand an. Kennst du da schon Beispiele oder jetzt auch gerade hier mit der Telekom MMS, was ihr da so umsetzt? Also was ist auch ein guter Weg, so eine readiness hinzubekommen?

[00:22:08.650] - Prof. Dr. Frank Schönefeld

Ja, also es gibt Beispiele, man findet die auf der Website der MMS, Referenzen. Da muss man vielleicht, sagen wir, unser Gespräch fokussiert sich ja stärker auf den generativen KI Teil. Das soll uns aber nicht blind machen für die Anwendung der analytischen oder diskriminativen KI. Dort können die Return of Invests auch wesentlich schneller gehoben werden. Deswegen sind viele Fälle, die wir dort reporten in den Referenzen, sind auch Anwendung analytischer KI, also AI Visioning, also intelligentes Betrachten von Eingangsformularen, von Eingangspaletten, aber auch von Schüttgütern etc. pp. Und darauf dann die Prozesse zu automatisieren. Trotzdem muss man sagen, dem Mittelstand fällt es in Gänze etwas schwerer, dort Tritt zu fassen. Wie gesagt, diese Abhängigkeit vom digital readiness Stand gibt es. Woran wir aber sehen, dass die Nachfrage und die Tendenz, dorthin zu gehen und sich dorthin zu bewegen, ungebrochen ist, sehen wir bei unserem Beratungsgeschäft. Wir haben dort eine Sparte, die KI Beratung natürlich in den Mittelpunkt stellt. Und die Kolleginnen und Kollegen hatten ein schweres Jahr, aber im Sinne von Last und Überlast und Nachfrage.

[00:23:32.690] - Prof. Dr. Frank Schönefeld

Also man merkt, da ist hoher Bedarf da und man geht jetzt die Schritte, dass man sich sachkundig macht und die die ersten haben natürlich die ersten Anwendungen im Feld und die weiteren werden folgen. Parallel dazu steigt natürlich die Reife der Produkte, auch die der Integrationsgrad der Produkte. Microsoft Copilot ist in aller Munde und das ist natürlich auch ein sehr einfacher Weg für einen Mittelständler, dann diese Experimente zu führen. Er muss natürlich bereit sein, den höheren Lizenzpreis dann aufzubringen. Das geht dem Mittelständler aber genauso wie dem Großunternehmen.

## Datenqualität, Halluzination und Kuratierung

[00:24:13.520] - Steffen Wenzel

Wenn wir an Microsoft Copilot denken gerade, du hast es erwähnt, gibt es auch Kritik. Also insbesondere was die verwendeten Daten anbelangt. Es wurden da auch viele Falschinformationen gefunden. Dinge wurden verwechselt, wo es dann auch teilweise wirklich zu vergleichen kam. Zu sagen, sucht lieber weiterhin bei Google, als euch auf diese generativen KIs zu verlassen ist. Sind das jetzt Kinderkrankheiten oder wie schätzt du das ein?

[00:24:39.390] - Prof. Dr. Frank Schönefeld

Na ja, zum Teil stecken sie in der Natur der Lösung. Das wissen wir ja. Aber es gibt inzwischen Gegenmittel, mit denen ich auch Halluzinationen sozusagen gegenchecken kann. Und wer mal aufmerksam so eine Gemini Antwort sich zum Beispiel anschaut, der wird dort zwei Farben feststellen. Und die eine Farbe wird verwendet. Ja, hier ist noch mal ein Gegencheck über eine Suchanfrage und das kann ich bestätigen. Das habe ich gefunden und das andere stammt rein aus dem statistischen Sprachmodell heraus. Das ist dann anfälliger für eine mögliche Halluzination. Ansonsten können wir sagen die, die Risiken und Qualität sind eigentlich die hat man sehr früh erkannt, welche alle existieren

und die sind auch alle noch gültig und insofern hat man aber auch zumindest die Theorie, wie ich sie behandeln kann, liegt eigentlich auch klar. Also wenn man es mal durchgehen: Halluzinationen hatte ich gesagt, ist inhärent wird man nie ganz weg kriegen. Ich kann aber Gegenchecks einbauen und damit reduziere ich das Risiko schon ein bisschen. Transparenz der Modelle: 405 Millionen Parameter bei LLaMA 3.3 oder gar 1,76 Trillionen bei GPT-4.

[00:26:03.270] - Prof. Dr. Frank Schönefeld

Das ist eine Größenordnung, die fällt uns schwer zu durchsteigen. Ich muss also Mechanismen einbauen, um die Nachvollziehbarkeit: Warum ist das so entstanden? Was ist dort erzeugt worden? Kann ich versuchen einzubauen, sogenannte explainable AI oder nachvollziehbar KI. Das wäre eine Variante. Und das Datenproblem ist natürlich allgegenwärtig. Da gilt die alte Gleichung shit in shit out. Also wer gut kuratiert, gut kuratierte Datensätze verwendet, wird von Anfang an das Niveau, das Ausgabenniveau seines Sprachmodells natürlich anheben. Kuratieren ist aber leicht gesagt, weil wir sprechen ja von ebenfalls 30 trillion, also das ist die amerikanische Trillion, Artefakten, die den Modellen zum Futtern gegeben werden. Und das zu kuratieren ist natürlich auch keine ganz einfache Aufgabe.

[00:27:01.330] - Steffen Wenzel

Wenn wir aber davon ausgehen, dass gewisse Datensätze schon quasi kontaminiert sind, also dass sie zum Beispiel diskriminieren, was ja ein großes Problem ist, ob es jetzt eine ethnische Diskriminierung, eine sexuelle oder eine Sonstige ist. Wie will man das kuratieren? Also das heißt, wie will man da sicherstellen, dass auch KI wirklich eine Weiterentwicklung ist? Das ist der erste Teil meiner Frage. Und der zweite ist, das ist ja auch ein Problem, wenn man sieht, wie politische Tendenzen gerade in den USA sich entwickeln. Klammer auf, Mark Zuckerbergs Proklamation jetzt in den letzten Tagen, Klammer zu. Kann man da nicht auch Angst bekommen, dass es hier gewisse Absichten geben wird, die genau gewisse Tendenzen verstärken will?

[00:27:46.720] - Prof. Dr. Frank Schönefeld

Es gibt widersprüchliche Entwicklungen, widersprüchliche Tendenzen in dem Umfeld. Wenn ich mich rein auf die wissenschaftliche Position zurückziehe, sage ich ich, ich kann es lösen, ich kann es durch Kuratieren lösen. Ob dieses Kuratieren gewollt wird oder ob, letztlich ist ja auch eine Kuratierung eine gewisse Anlegen von Maßstäben. Und da steckt natürlich auch schon eine Bias Gefahr sozusagen wieder drin. So ist ja auch die Argumentation von Musk und Co, warum man das nicht zulassen sollte. Also wir hätten technische Möglichkeiten dort vorzubeugen. Aber es kommt natürlich auch die organisatorische, die menschliche, die politische Komponente hinzu. Und die kann nicht vollständig technologisch abgewendet, abgebildet werden.

## Europäische Entwicklungen und Regulierungen

[00:28:39.790] - Steffen Wenzel

Glaubst du, dass wir in Europa da einen besseren Ansatz fahren mit dem KI-Act und natürlich auch den mit der Regulierungswut in Anführungszeichen, die ja jetzt gerade auch Musk und Zuckerberg uns vorwerfen. Oder glaubst du das ist zu stark?

[00:28:56.350] - Prof. Dr. Frank Schönefeld

Ja, ich, ich möchte da mal den Chef des norwegischen Staatsfonds zitieren, Der heißt Nikolai Tanger. Das ist der größte Fonds der Welt, wenn man so will. Dort sind praktisch 1,5 % aller weltweiten Unternehmen, gehört diesem norwegischen Staatsfonds. Er hat gesagt: "Europa hat viel Regulierung und wenig KI und die USA hat viel KI und wenig Regulierung." Und bei ihm schwingt da durchaus auch die Wertung, so wie ich sie jetzt zitiert habe, natürlich mit. Er sieht also durchaus die Gefahr, dass man überreguliert ist und dass man dadurch sich Möglichkeiten des Fortschritts, der Innovation sozusagen vergibt oder zumindest verlangsamt. Und wenn man jetzt in die europäische ActGeschichte schaut, dann stellt man natürlich auch fest: "Oh Gott, da gibt es aber eine Menge.": das ist der Data Governance Act, das ist der Data Act,

das ist der AI Act, das ist der Digital Marcets Act. Ja, das ist das GbtR, also das Data Privacy Act könnten wir es auch nennen, das heißt, es ist auch sehr allumfassend, im besten Wollen und mit den besten Zielen natürlich versehen.

[00:30:20.380] - Prof. Dr. Frank Schönefeld

Aber ich, wir erleben es ja jetzt schon, dass gewisse Innovationen gar nicht mehr nach Europa ausgerollt werden oder nicht sofort ausgerollt werden, sondern mit Verzögerung ausgerollt werden. Und das heißt doch auch, dass man einen Großteil der Bevölkerung der Nutzungsmöglichkeiten praktisch von den bahnbrechenden Innovationen abschneidet. Und das ist natürlich eine Gefahr. Kann man deswegen sagen, wir brauchen keine Regulierung? Nein, das würde das Kind mit dem Bade ausschütten. Das heißt es nicht. Aber ein Augenmaß ist sicherlich angemessen an der Stelle. Und wenn ich noch mal einen Punkt sagen kann: Wenn das der AI-Act ist ja eine risikobasierte Betrachtung. Damit fängt es schon an, wenn ich KI sozusagen in seiner Regulierung als Risiko betrachte, dann fehlt da das Chancenbild dazu. Da fehlt, dass der Produktivitätsgewinn dazu und da ist von vornherein sozusagen eine einseitige Betrachtung drin und vor der warne ich auch. Aber Regulierung muss sein. Zum Beispiel denke ich, was Australien gerade macht, keine sozialen Medien, bis jemand 16 ist. Das ist auch eine Art der Regulierung und die wird viel mehr erreichen, als welcher Act auch immer.

[00:31:42.400] - Steffen Wenzel

Ist ja auch ein Act. Wird ja auch übrigens hier in Deutschland ja auch diskutiert, in vielen Familien auch ist das gerade natürlich immer wieder ein Punkt und ich finde es auch sehr spannend, darüber zu reden, trotz aller Regulierungsversuche und ich möchte es gar nicht Wut nennen, aber es ist einem immer im Munde, gibt es ja auch in Deutschland ein paar Unternehmen oder zumindest eins, was ich jetzt dort auch positioniert hat: Aleph Alpha. Kannst du dazu ein bisschen was sagen? Wie schätzt du die ein?

[00:32:08.240] - Prof. Dr. Frank Schönefeld

Ja, ich glaube, in letzter Zeit ist es ein bisschen ruhiger um sie geworden. Aber wir reden ja auch über 24, und da waren Sie schon noch in aller Munde. Ich glaube, Sie versuchen, Ihr Geschäftsmodell gerade ein bisschen zu readjustieren. Ganz am Anfang haben Sie versucht, mit dem Sprachmodell selber mit sozusagen die Schlagzeilen zu bestimmen. Jetzt sind Sie eher in so einem gemischten Servicemodell und Produktmodell unterwegs. Aus meiner Sicht bleibt abzuwarten, wo Sie sich in einem Jahr wiederfinden. Aber generell ist das Thema noch mal dankbar und spannend. Was ist in Deutschland passiert? Technologisch? Was ist in Europa passiert? Und den Herrn Tanger hatte ich ja schon zitiert. Aber trotzdem brauchen wir, es ist nicht so, dass totaler Stillstand herrschen würde, im Gegenteil. Es gibt eine einzelne Entwicklung, die sehr spannend sind, die Mut machen. Ich würde da die Black Forest Labs dazu zählen. Die haben ja eine Geschichte über Stable diffusion hin zu den Black Forest Labs, sind jetzt natürlich auch als Start up in den USA präsent, haben aber auch die Homebase im Breisgau in Freiburg und sind, was die Schnelligkeit einer generativen KI für Bildgenerierung betrifft, es ist das Flux.1 oder Flux1.1 sind sie unerreicht, das heißt dort wird im Moment das Topniveau durch eine deutsche Firma, man sicherlich auch international natürlich inzwischen basiert, bestimmt. Das ist gut zu wissen. Und an zweiter Stelle ist sicherlich Mistral zu nennen. Die Franzosen, die in der Leistungsfähigkeit mithalten können, die liegen, wenn man das rein auf europäische Sprachen beschränkt, liegen die ungefähr auf Platz zwei, Platz drei der Leistungsfähigkeit. Das ist gut. Und haben auch einen kompletten Modell-Zoo. Inzwischen auch von klein nach groß, von Textgenerierung zu Bildgenerierung, praktisch alles decken sie ab. Also der europäische Marktführer für große Sprachmodelle. Und ganz spannend im Dezember hat auch das OpenGPT-X Project ein erstes Resultat geliefert namens Teuken, Teuken-7B. Die sieben, natürlich bleibt sieben, das B sind die Milliarden, also ein 7 Milliarden Parameter Modell und hat sich in der ersten Benchmarkrunde auf Platz 25 etwa einsortiert, für europäische Sprachen. Heute Morgen habe ich noch mal nachgeschaut, da waren sie auf Platz 17. Also das ist gut so und da kann ich nur wünschen und auch appellieren, dass man den Prozess aufrechterhält und sich auch graduell verbessert. Weil aus meiner Sicht muss es möglich sein, auch da in die Spitzengruppe vorzustoßen.

[00:35:10.240] - Prof. Dr. Frank Schönefeld

Wenn man sich die Benchmarks genau anguckt, stellt man fest: Teuken ist gar nicht so schlecht, aber rechnen kann es nicht. Also sobald man da den GSM8K Benchmark ein bisschen nach oben zieht von 0,11 im Moment auf vielleicht 05 oder 06, dann macht man gleich einen Sprung um zehn Plätze im gesamt Benchmark. Das sollte möglich sein aus meiner Sicht.

## Geschäftsmodelle

[00:35:32.170] - Steffen Wenzel

Du hast jetzt eine ganze Menge an Sprachmodellen und generativer KI Modellen genannt. Wo findet man die denn eigentlich so im Einsatz? Also man kennt natürlich immer ChatGPT, man kennt Copilot usw. bei Bing, aber wo findet man so die anderen?

[00:35:47.770] - Prof. Dr. Frank Schönefeld

Also eine gute Quelle sind natürlich die KI Blogs der Hersteller selber. Die lese ich auch regelmäßig und da merkt man natürlich, welche Facetten das weiterhin hat. Und die Benchmarks, da gibt es spezialisierte Seiten für die offenen Modelle, zum Beispiel auf Hugging Face. Da gibt es das berühmte Leadership Board von Hugging Face und dort werden die fünf Standard Benchmarks unterzogen und nach dem Benchmarks gerankt. Es gibt aber auch darüber hinaus andere Benchmarks, auch gute Benchmarks Seiten. Da kommen noch stärker dann die Kosten rein. Also welches Modell ist wie teuer? Und da bildet sich auch gerade eine neue Bewertungssystematik raus. Man sagt einfach was kosten 1 Million Token im Inputstream und was kostet 1 Million Token im OutputStream? Und da hat man dann auch einen Kostenvergleich der Modelle, zumindest auf API Basis. Und auch da liegen durchaus dann Größenordnungen dazwischen. Also ein LLaMA wirbt dort mit Kosten von 0,10 € für 1 Million Begriffe im InputStream und 0,40 € im Outputstream. Also das ist dann schon wettbewerbsfähig.

[00:37:08.640] - Steffen Wenzel

Ich meinte auch stärker: Wo werden die eingesetzt? Also haben die ja auch einen ChatGPT hat ja ein Abo-Modell als Businessmodell. Also was, was ist deren Businessmodell?

[00:37:18.540] - Prof. Dr. Frank Schönefeld

Also zu Geschäftsmodellen können wir natürlich noch mal kommen. Über Einsätze hatten wir vorhin schon gesprochen, letztlich im Privatbereich sehr sinnvolle Applikationen und im Geschäftsbereich ist letztlich kein Geschäftsprozess außen vor. Hatten wir auch schon diskutiert. Aber jetzt zur Frage: Welche Geschäftsmodelle haben die Anbieter selber? Und da nehmen wir noch keine großartige Weiterentwicklung gegenüber den klassischen digitalen Geschäftsmodellen wahr. Natürlich die Paid Plan Modelle, also Bezahlplanmodelle. Das O1 kostet im Moment, also das fortgeschrittenste GPT Modell, 200 $ im Monat. Und wenn ich diese 200 $ zahle, da habe ich inzwischen natürlich auch die Videogenerierung mit drin. Da ist Sora mit drin und Sora ist im Moment das einzige verfügbare Videogenierungssystem in einem Paid Plan. Die anderen haben die Lösungen auch. Also Meta hat auch eine tolle Videogenerierung, hat es aber noch nicht freigegeben und die anderen stehen wahrscheinlich kurz davor, haben es aber noch nicht freigegeben.

[00:38:29.490] - Steffen Wenzel

Ganz kurz Videogenerierung heißt ich kann per Textbasis etwas eingeben, eine Szene, die ich irgendwie brauche. Ob es jetzt ein Imagevideo für eine Company ist oder etwas anderes und dann wird das automatisch durch die KI generiert?

[00:38:44.490] - Prof. Dr. Frank Schönefeld

Das ist es.

[00:38:45.030] - Steffen Wenzel

Ich kann Lichteinstellungen wählen? Alles mögliche?

[00:38:47.760] - Prof. Dr. Frank Schönefeld

Alles, alles möglich. Und die, die Beispiele sind auf den KI-Blogs abrufbar und sind sehr impressiv, natürlich. Weil ich natürlich meiner Fantasie freien Raum lassen kann, dass dort die Gefahr von Fakevideos natürlich noch mal um ein Erhebliches steigt. Jetzt mal außen vor gelassen, aber wir waren von den Geschäftsmodellen gekommen. Also wir haben nach wie vor die Paid Plans. Dann haben wir aber im Gegensatz dazu eher die Plattformmodelle, wie sie vielleicht bei Meta selber anzutreffen sind oder wie sie bei Apple anzutreffen sind. Dort ist eher die Idee: Ja, ich muss die Nutzer auf meiner Plattform auf meine Plattform ziehen und ich muss sie auf meiner Plattform halten. Und dafür ist KI-Funktionalität unverzichtbar. Ja, und dann habe ich sie auf der Plattform und von dort aus findet häufig dann so ein Übergang in so ein Freemium Modell statt. Das heißt, die ersten zehn Videos, die ersten zehn Bilder sind frei zu generieren oder die ersten 20 Anfragen sind frei zu generieren, und wenn ich das übersteigen will, weil ich es eben sehr häufig nutzen möchte, dann kommt langsam so ein Paid-Modell dazu.

[00:40:01.760] - Prof. Dr. Frank Schönefeld

Was wir noch nicht sehen ist, wie sich Suche, Suche wird letztlich durch Werbung finanziert, das ist das Geschäftsmodell von Google, wie so eine intelligente Suche, wie sie jetzt auch bei Gemini angeboten wird, wie die sich refinanziert, das ist noch ein bisschen offen. Natürlich hat man da auch Ansätze zur Werbung, aber es ist schwieriger zu platzieren, weil es tiefgründiger in der Antwort sein muss. Und ich denke, auch Perplexity, die auch so eine intelligente Suchmaschine ist, hat auch noch keine richtige Antwort gefunden. Aber auch hier ist die Idee: erst mal Masse anziehen, und habe ich Masse auf der Plattform, dann wird mir auch schon was zum Geschäftsmodell einfallen.

## Spezialisierung von Sprachmodellen

[00:40:51.770] - Steffen Wenzel

Was wir letztes Jahr ein bisschen prognostiziert hatten oder bzw. besprochen hatten, war, dass wir sicherlich eine Ausdifferenzierung der einzelnen Sprachmodelle nach Fachthemen erwarten, also juristische Themen, Marketingthemen, die dann in Unternehmen oder auf ausgerichtet auf unterschiedliche Unternehmen dann auch so produziert werden. Hat das stattgefunden?

[00:41:14.870] - Prof. Dr. Frank Schönefeld

Also diese Spezialisierung gibt es. Wir hatten ganz am Anfang haben wir das Thema Edge-LLMs diskutiert. Das ist natürlich schon eine Spezialisierung. Und nach wie vor ist natürlich auch in großer Nachfrage: Wie kann ich unternehmensinternes Wissen mit der Intelligenz eines Sprachmodells verbinden? Und dort gibt es verschiedene Methoden, wie mir das gelingt. Es gibt inzwischen auch gute Produkte, wo das quasi nahtlos geschieht. Also Gemini NotebookLM ist so ein Beispiel für so ein Produkt. Als Technologie steckt meistens dieses RAG Verfahren dahinter. Retrieval Augmented Generation, das heißt, ich lege neben das Sprachmodell lege ich noch eine Dokumentenbasis, eine klassische Datenbasis. Und ich verbinde diese klassische Datenbasis intelligent mit den Fähigkeiten des Sprachmodells. Also insofern ja, diese Spezialisierung gibt es, wenn man Mistral noch mal hernimmt. Dort kann ich praktisch die Spezialisierung: Generierung für Text, Generierung für Bilder, Generierung für Code und das auch in verschiedenen Größen sozusagen abstufen. Und dann kann ich eben je nach Einsatzzweck kann ich sagen, ich ich brauche ja nur die Codeunterstützung, also die Programmiercode-Unterstützung. Das reicht mir, dann brauche ich, da brauche ich keine Bildergenerierung drin, aber die vollen Modelle, die sogenannten Multimodalen Modelle, die eben auch als Input ein Bild verstehen. Das sind natürlich dann die die Großen, das 2.0, das haiku, das sonnet, das sonnet, und so weiter und sofort.

[00:43:06.740] - Steffen Wenzel

Für mich ist das jetzt so eher auch ein Thema, gerade in so relativ eingesessenen Fachgebieten wie zum Beispiel das Ganze, also ganze Juragebiet, also alles, was mit Rechtsprechung zu tun hat, weil mein Schwager ist beispielsweise Rechtsanwalt, der hat gesagt, also mit ChatGPT, da komme ich nicht viel weiter, selbst mit der Pro-Version nicht, also mit der kostenpflichtigen Version nicht. Ich weiß, dass es ja viele Firmen gibt, die jetzt quasi Lawtech machen und natürlich da auch KI einsetzen. Warum sind die noch nicht so weit? Meines Wissens gibt es da noch nicht richtig was auf dem Markt, wo man sagen kann, ja, da kann man was wirklich Neues erwarten.

[00:43:45.830] - Prof. Dr. Frank Schönefeld

Das muss man auch noch mal systematisch Untersuchen. Ich ich glaube, im Moment ist es noch geprägt durch Einzelwahrnehmungen, Einzelschilderungen und so wie hier das Beispiel ja, ich kann es doch noch nicht so richtig verwenden, habe ich Beispiele von einem Professorenkollegen an der HTW, der sagt, ich nutze es nur und er ist Jurist und lehrt in diesem Umfeld und macht praktisch sämtliche Fallvorbereitungen praktisch über über generative KI. Und dieser ursprüngliche Test, da gibt es das sogenannte Uniform Bar Exam in den USA, das muss jeder, der zum Rechtsanwalt zugelassen werden will, bestehen. Eine typische Durchfallquote für Menschen ist 40%. Ja und das ChatGPT liegt dort bei 88-90 % im Erfolgsfall aller Fragen. Aber wie gesagt, auch das sind Einzelbeobachtungen, Einzelwahrnehmungen. Ich denke, die systematische und wirklich, kann ich es produktiv in meinem dem Prozess, den ich typischerweise mache im Arbeitsprozess, kann ich dafür einsetzen, das liegt noch nicht für alle Prozesse vor. Aber insofern ist die Juristerei oder die die Rechtsbeispiele steht eigentlich exemplarisch für alle Arbeitsprozesse. Auch unsere Entwickler und Architekten fragen sich natürlich: Wie weit ist es schon zuverlässig nutzbar, wie hoch ist mein Arbeitsaufwand?

[00:45:28.060] - Prof. Dr. Frank Schönefeld

Aber die Tendenz und ich glaube, das sollte uns klar sein, die Tendenz ist ganz klar: Der KI Anteil wird steigen und der wird auch qualitativ so hochwertig werden, dass er genutzt werden kann. Ich bin unlängst gefragt worden: "Kann man die die nächste Etappe irgendwie mit einer Überschrift versehen, die wir da vor uns sehen." Und ich habe, meine Antwort wäre: "Wir werden ein Jahrzehnt der intensiven Kollaboration zwischen Mensch und Maschine, Mensch und KI erleben."

## Transparenzpflicht?

[00:46:01.150] - Steffen Wenzel

Das können wir gleich noch mal vertiefen. Zum Abschluss wollen wir natürlich noch mal hinschauen: Was passiert jetzt in den nächsten Jahren? Im nächsten Jahr? Damit wir es auch wieder bei un serem nächsten Treffen verifizieren können, ob es wirklich passiert ist. Ich hätte noch eine Frage, noch mal kurz vorgeschoben, zum Thema Transparenz, weil du jetzt auch eben noch mal die Universität genannt hast. Wenn dein Professorenkollege dort quasi Klausuren oder ähnliches vorbereitet: Macht er das transparent, dass er das tut gegenüber seinen Studierenden? Und wie geht ihr damit um, dass die Studierenden das selbst einsetzen?

[00:46:35.020] - Prof. Dr. Frank Schönefeld

Ja, macht er. Er macht es Transparent. Halte ich auch für eine wichtige Regel. Auch das kann man natürlich regulieren, dass es diese Transparenzpflicht gibt. Von Studenten her ist das schwieriger einzufordern. Deswegen merkt man ja auch, dass die Art und Weise, wie man Wissen erfragt, wie man Wissen abtestet, dass die sich auch verändert und stärker auch individuell wird, stärker, persönlicher wird. Das heißt, gewisse Verfahren stehen uns nicht mehr zur Verfügung. Da hat die KI inzwischen andere Gesetze erzeugt. Ja, was das betrifft, bin ich ganz strenger Anhänger einer Transparenzpflicht.

[00:47:16.200] - Steffen Wenzel

Und wie machen wir das mit der Transparenz, dass KI irgendwo eingesetzt wird? Wird ja auch viel über Wasserzeichen usw. diskutiert. Kommen wir da voran?

[00:47:24.660] - Prof. Dr. Frank Schönefeld

Auch hier gibt es die Antwort, die ich vorhin schon gegeben habe. Technologisch geht alles, und die die Wasserzeichen in generierten Bildern zum Beispiel oder generierten Videos oder generierten Texten, auch das kann man verstecken. Steganographie wäre eine Möglichkeit. Es muss halt mit organisatorischen und politischen Vorgaben und Wünschen ebenfalls in Übereinstimmung gebracht werden. Und das ist schwer zu prognostizieren, wohin es da geht.

## Ausblick: Roboter als Individuum

[00:47:52.290] - Steffen Wenzel

Dann gucken wir mal, was vielleicht besser zu prognostizieren ist. Du hast schon gesagt, ein Jahrzehnt des Menschen und der Maschine, glaube ich, wird auf uns zukommen. Was erwartest du jetzt erst mal für 2025? Und vielleicht ist das viel zu kurzfristig gedacht, vielleicht auch wirklich für die nächsten Jahre.

[00:48:07.770] - Prof. Dr. Frank Schönefeld

Ja, ich hatte schon erwähnt, dass das Energieproblem ein Drängendes ist. Das macht es ja auch teuer. Und eine Lösung, die den Energieverbrauch bei gleichbleibender Intelligenz gleich hoher Intelligenz reduzieren würde, wäre natürlich sehr wünschenswert. Es gibt auch schon ein paar Überlegungen, wie das gelingen könnte. Man kann gewisse Teile des Algorithmus verändern. Der sogenannte Attention Mechanismus, der die Zusammenhänge zwischen den Eingabewörtern zueinander untersucht, der skaliert leider quadratisch. Je länger mein Input ist, desto höher und größer ist der Aufwand, diese Zusammenhänge zu berechnen. Und dort gibt es inzwischen andere Algorithmen: Mamba, Mamba-2, die praktisch diese Skalierung linear halten. Und das gleiche könnten wir noch mal versuchen in den Schichten des Transformers. Dass dort einfach, ja eine stärkere, vielleicht Erinnerung, was schon verarbeitet wurde, eingebaut werden kann und dass damit eine signifikante Reduktion des Energieverbrauchs erfolgen kann. Um das noch mal ins Bild zu rücken: Das menschliche Gehirn braucht etwa 20 Watt pro Stunde und so ein KI-Mechanismus für, für eine Frage braucht etwa 85.000 Watt in der Stunde. Das ist also Faktor 4000. Auch das noch.

[00:49:41.000] - Prof. Dr. Frank Schönefeld

Wenn wir jetzt die Maschine auf 20 Watt beschränken würden, dann wäre es aber mit der Intelligenz nicht mehr weit her. Müssen wir einfach mal so sagen. Also auch das muss man in Perspektive praktisch zueinander bringen. Und deswegen denke ich, ist das natürlich eines der ganz wesentlichen Themen. Die zweite große Stoßrichtung ist natürlich die: Wie intelligent ist es denn nun? Also, diese Leistungsfähigkeit und diese Vergleichbarkeit mit menschlicher Intelligenz noch einen Tick zu erhöhen. Und da gibt es interessante Aussagen dazu. Der Chefwissenschaftler von Meta ist ein Franzose, Yann LeCun, einer der Großen im Gebiet, und der hat gesagt: Na ja, was erwartest du denn von einem System, was 30 Trillionen Begriffe gelesen hat? Du wirst niemals so gut sein wie das System, weil du nicht 30 Trillionen Begriffe lesen kannst. Das wirst du nicht schaffen. Aber dieses System hat keine einzige Bewegung in der realen Welt gemacht, hat kein Tastgefühl, kein Tastfeedback zurückbekommen, hat im Prinzip nichts gesehen etc. pp. Das heißt, ein Blick ins Buch und zwei ins Leben. Die zwei Blicke ins Leben fehlen dem System noch.

[00:50:56.360] - Prof. Dr. Frank Schönefeld

Und deswegen sagt er: "Wir müssen, wenn wir richtige Intelligenz oder eine neue Stufe an Intelligenz erzeugen müssen, dann müssen wir natürlich auch die Inputreize einfach verbreitern, müssen wir steigern, müssen ein real world modell erzeugen, und dann werden wir natürlich auch real world intelligence am langen Ende sehen." Die reinen Mechanismen, wie man trainiert und wie man das

abbildet und in neuronalen Netzen ablegen kann, speichern kann, die scheinen allgemeingültig genug zu sein. Das heißt, der Mechanismus trägt uns noch eine ganze Zeit. Aber ja, wir wir sind natürlich noch ganz am Anfang, Aber mit den multimodalen Trainings merkt man ja, das ist ja nicht nur Text, das sind ja auch schon Bilder. Und jetzt gehen die ersten Systeme, insbesondere Nvidia, auch dazu über, stundenlanges Videomaterial in solche real world simulatoren zu füttern. Und das wird, glaube ich, die nächste Stufe sein an Intelligenz. Eine Vielzahl von von Reizen, nicht nur Text-Reize, sondern real world Reize werden den Systemen zur Kenntnis gegeben und dann werden wir sehen, was dort für neue Formen von Intelligenz entstehen können.

[00:52:12.400] - Steffen Wenzel

Aber es ist immer ein Spüren des anderen oder der anderen. Also wenn ich, die Maschine selbst wird ja erst mal nichts spüren. Oder müssen wir auch vielleicht in eine haptische Vorstellung gehen, dass es eben nicht mehr irgendwelche Geräte sind, die in Rechenzentren sind, sondern wirklich das, was man ja immer in solchen Science Fiction Filmen sieht, dann wirklich Maschinen, die dann selbst auch riechen, anfassen, spüren können?

[00:52:36.740] - Prof. Dr. Frank Schönefeld

Auch da gibt es erste wissenschaftliche Untersuchungen dazu. Was ist notwendig, damit solche Effekte entstehen können? Und die mind needs body These, also reales Bewusstsein braucht auch einen abgeschlossenen Körper, so ungefähr ist dort das Postulat, die zählt natürlich dazu. Aber auch da sind wir nicht so weit weg. Die robotrevolution schreitet ja auch ungebremst voran. Und in so einer Roboterkonstellation kann natürlich auch ein ganz anderes Feedback, noch mal. Ein Roboter ist dann auch ein Individuum, muss man einfach so sagen, ja. Und da können noch mal völlig neue Zusammenhänge und Feedbackschleifen entstehen.

## Verabschiedung

[00:53:20.030] - Steffen Wenzel

Frank, zum Abschluss: Was würdest du dir noch wünschen für 2025? Also jetzt ein wirklicher Wunsch, wo du sagst zu diesem Thema KI, das wäre gut, wenn das passieren würde. Ist nur einer, das ist dann schwer.

[00:53:35.610] - Prof. Dr. Frank Schönefeld

Ja. Ich hoffe, dass die positiven Seiten der KI, die produktivitätssteigernden Seiten, aber auch die Menschen selbst bereichernden Seiten, immer im Vordergrund bleiben und größer bleiben, als die unzweifelhaft vorhandenen Risiken.

[00:53:52.320] - Steffen Wenzel

Das ist doch ein sehr gutes Schlusswort. Frank, vielen Dank, dass du heute hier zu Gast warst und ich freue mich jetzt schon auf das nächste Jahr und ich wünsche dir eine gute Zeit bis dahin.

[00:54:01.350] - Prof. Dr. Frank Schönefeld

Danke dir. Tschüss.

[00:54:03.210] - Steffen Wenzel

Ja, und auch vielen Dank, dass Sie heute hier zu Gast waren und uns zugehört haben. Ich hoffe, es war eine kurzweilige Zeit für Sie. Und wenn Sie weitere Informationen benötigen, finden Sie die wie immer in den Shownotes. Und Sie können uns natürlich auch gerne auf den bekannten Kanälen abonnieren. Das würde uns auch sehr freuen. Und ansonsten wünschen wir Ihnen eine gute Zeit. Bis dahin alles Liebe