c't KI-WISSEN

Kritische Analysen & kreative Praxis

Monopolkonzerne aus den USA

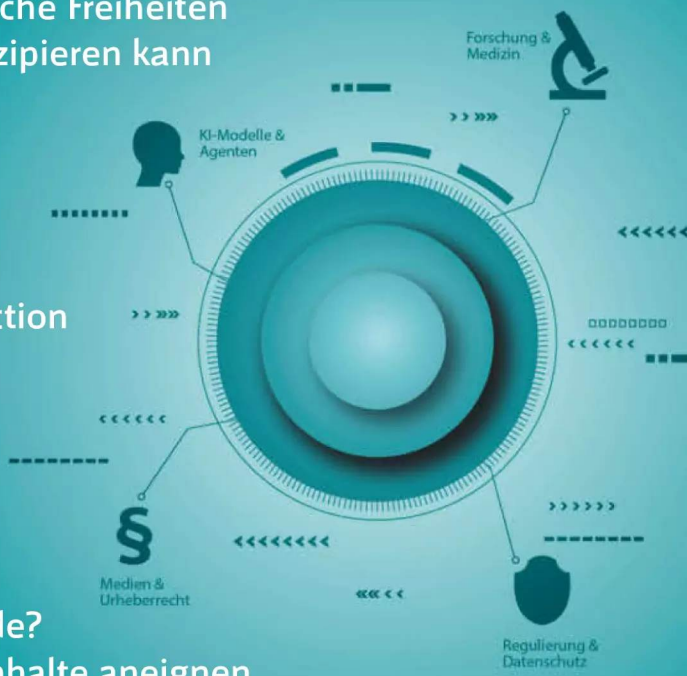
Auswirkung auf Wirtschaft, Umwelt und bürgerliche Freiheiten
Suche nach Alternativen: Wie Europa sich emanzipieren kann

Blick in die Zukunft

Smart Glasses zwischen Realität und Science-Fiction
Krebserkennung und Psychotherapie mit KI

Kultur & Urheberrecht

Song-Generatoren: Stehen Musiker vor dem Ende?
Diebstahl wird Eigentum: Wie KI-Anbieter sich Inhalte aneignen



Im Netz der KI-Agenten

Wie Sprachmodelle Apps und Internetdienste steuern
Grundlagen lernen von RAG bis MCP und eigene Server aufsetzen

€ 14,90
CH CHF 27.90
AT € 16,40
LUX € 17,10

04
4 196982 314900

Editorial

Liebe Leserinnen und Leser,

der französische Philosoph Pierre-Joseph Proudhon prägte einst den geflügelten Satz „Eigentum ist Diebstahl.“ Karl Marx stellte ihn in einer bissigen Kritik vom Kopf auf die Füße. Angesichts der riesigen Datensammlungen großer Konzerne für ihre KI-Modelle fragen heute immer mehr betroffene Autoren und Kulturschaffende: Wird da nicht Diebstahl zum Eigentum?

Besonders Musiker sind davon betroffen, wie wir auf den Seiten 120 und 138 zeigen. Im Kleingedruckten ihrer AGBs nehmen sich US-Firmen wie Suno das Recht heraus, sämtliche hochgeladenen Nutzerdaten nach eigenem Gutdünken zu verwenden. Diese unschöne Sitte greift auch bei anderen KI-Anbietern immer mehr um sich. Das führt direkt zum exponentiellen Wachstum der großen KI-Modelle. Welche Rolle sie für die sieben größten Konzerne aus den USA spielen, die als Gatekeeper großer Plattformen inzwischen eine so große wirtschaftliche Macht ausüben, dass sie die Regeln des freien Marktes aushebeln, beleuchten wir ab Seite 158.

Am Beispiel der Schweiz schauen wir ab Seite 176 und 184, wie es anders gehen kann, nämlich mit echten Open-Source-Modellen, deren Algorithmen für jeden einsehbar sind, und transparenten Trainingsdaten, aus denen sich Inhalte entfernen lassen, wenn deren Urheber damit nicht einverstanden sind. In der Praxis sehen wir einen solch offenen Umgang leider noch viel zu selten. Deshalb erklären wir zum Einstieg die neuesten Entwicklungen der KI und vergleichen Modelle miteinander. Entwickler können damit ihre eigenen Server aufsetzen und KI-Modelle autonom betreiben (Seite 58). So lernen Sie die Bedeutung der Reasoning-Modelle (ab Seite 6), der Retrieval-Augmented Generation (RAG ab Seite 26) und des Model Context Protocol (MCP ab Seite 52) kennen, mit denen Forscher Halluzinationen verringern, den Ressourcenbedarf senken, Spezialwissen einbinden und der KI ermöglichen, komplexe Dienstleistungen und Tools im Internet zu nutzen.

Was die Zukunft bringt, kann man sich schon heute durch die neue Generation der Smart-Glasses ab Seite 78 ansehen. Sie blenden im Alltag Informationen zur Umwelt ein – ähnlich wie moderne Autos bereits Wegweiser auf die Windschutzscheibe projizieren. Dieses Heft können Sie freilich noch ohne solche Überwachungssensoren lesen.

Dabei wünsche ich Ihnen viel Spaß,

Hartmut Gieselmann

Inhalt

KI-MODELLE & AGENTEN

Die Entwicklung ist dieses Jahr rasant weitergegangen: Von Reasoning-Modellen mit Deepseek & Co. über neue Methoden, mit denen man Modellen Fachwissen beibringt, bis hin zu Agenten, die Webdienste über das Model Context Protocol bedienen. Wir erklären die Grundlagen und geben Tipps zum praktischen Einsatz.

FORSCHUNG & MEDIZIN

Zwar ist die Medizin noch nicht so weit wie bei Pille vom Raumschiff Enterprise, aber KI unterstützt bereits bei manchen Diagnosen. Problematisch wird es jedoch, wenn ChatGPT die Psychotherapie übernimmt oder die Modelle verstorbene Liebste simulieren sollen. Smart Glasses erleben derweil ihren zweiten Frühling.

MEDIEN & URHEBERRECHT

Künstler sind von den neuen KI-Generatoren besonders betroffen, denn viele Firmen greifen sich ihr Material, trainieren Modelle und kopieren ihren Stil. Bei Musik-Generatoren ist der Einsatz juristisch besonders heikel. Unproblematischer ist die KI-Hilfe beim Verwalten von Fotos und Videos.

REGULIERUNG & DATENSCHUTZ

KI-Modelle verhelfen den großen US-Firmen zu mehr Wachstum und zementieren ihre Monopolstellung. Eine Alternative fanden wir bei echten Open-Source-Modellen aus der Schweiz. Umstritten ist hingegen der polizeiliche Einsatz der Überwachungssoftware von Palantir. Hinkt die Regulierung in der EU hinterher?

ZUM HEFT

Bild: Rudolf A. Blaha

Acht Sprachmodelle mit Reasoning

Eine neue Generation von Sprachmodellen soll komplexe Aufgaben genauer lösen als bisher. Dazu entwickeln sie Schritt für Schritt Lösungsstrategien. In unserem Test müssen chinesische Modelle zeigen, bei welchen Aufgaben sie brillieren und ob sie die KI-Großmacht USA herausfordern können.

Von **Jo Bager und Hartmut Gieselmann**

Sprachmodelle mit der Fähigkeit, zu argumentieren und ihre Handlungen zu begründen, sind faszinierend: Fast könnte man meinen, ihnen beim „Denken“ zuzusehen, wenn man beobachtet, wie sie Aufgaben in einzelne Schritte aufteilen. Das erklärt den Ruck, der durch die Branche ging, als das chinesische Unternehmen DeepSeek im Januar sein Modell R1 für jedermann frei zugänglich bereitstellte. In puncto Leistungsfähigkeit konnte es mit wesentlich teureren Modellen konkurrieren. Dabei kostete seine Entwicklung nach Angaben von DeepSeek einen Bruchteil dessen, was ChatGPT, Gemini und andere bekannte US-Modelle verschlingen.

Doch DeepSeek ist nicht allein: Andere chinesische Anbieter wie Alibaba und Tencent legen ebenfalls Modelle mit Reasoning-Fähigkeiten nach. Wegen des US-Embargos haben chinesische Entwickler weniger Ressourcen zur Verfügung und das zwingt sie dazu, besonders sparsam damit umzu-

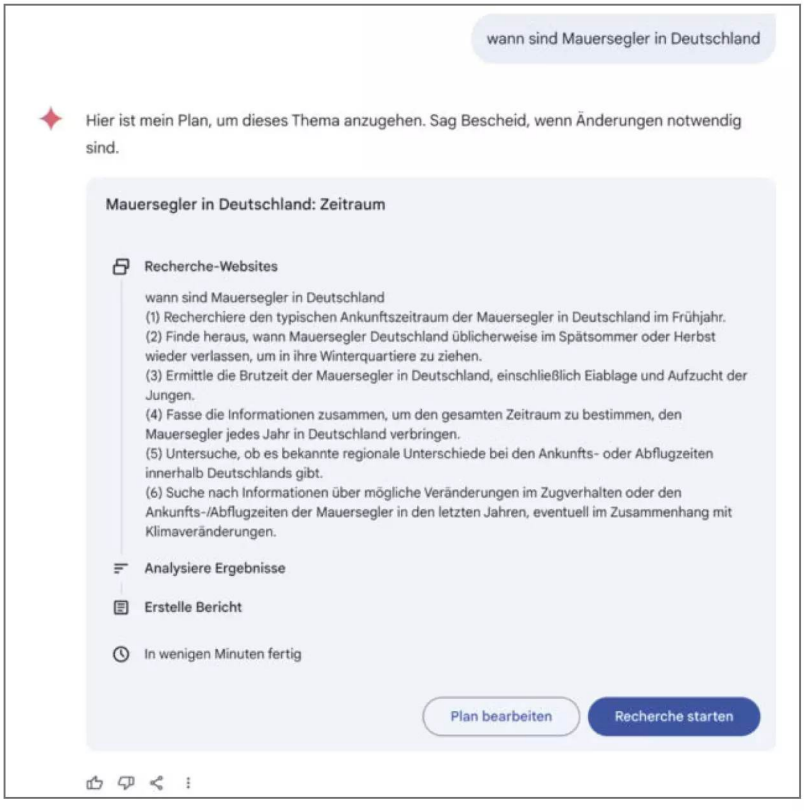
gehen. Anbieter und Kunden in Europa könnten von der Konkurrenz profitieren. Denn im Unterschied zu den meisten US-Modellen kann man die chinesischen Modelle frei herunterladen und oft mit deutlich geringeren Hardware- und Speicheranforderungen selbst betreiben.

In diesem im April 2025 durchgeführten Test stellen wir die neuesten und wichtigsten Reasoning-Modelle der US-Firmen Google, OpenAI, Perplexity und xAI den chinesischen Angeboten von Alibaba, DeepSeek und Tencent gegenüber, die auch in Deutschland als Dienst oder App verfügbar sind. Im Vergleich dazu muss sich OLMo von Microsoft-Gründer Paul Allen beweisen, das derzeit als das beste Open-Source-Modell gilt. Allens Firma Ai2 gibt nicht nur das fertig trainierte Modell kostenlos ab, sondern dokumentiert auch sämtliche Details zum Training und den kompletten Trainingskorpus. Damit ist es eines der wenigen LLMs, die tiefe Einblicke in ihre Funktionsweise gestatten; alle anderen sind proprietär oder halten zumindest die Trainingsdaten geheim.

Planen, abarbeiten, erklären

Doch was bedeutet Reasoning eigentlich genau? Der Begriff steht für eine Reihe von Fertigkeiten, die Fehler und Halluzinationen verringern sollen. Wenn ein Nutzer den Modellen eine Frage oder Aufgabe stellt, schreiben sie nicht sofort die Lösung auf, sondern formulieren zunächst einen Lösungsweg und klopfen diesen Schritt für Schritt ab. Wenn sie dabei in eine Sackgasse geraten, können sie auch wieder umkehren und eine andere Strategie wählen.

Um dies zu ermöglichen, wurden die Modelle nicht wie bisher nur mit vorgefertigten Frage-Antwort-Paaren trainiert und durch menschliches Feedback belohnt. Ihr Training wurde um sogenanntes Reinforcement Learning (RL) erweitert, das die Modelle dazu animiert, viele verschiedene Lösungswege auszuprobieren, die ein automatischer Schiedsrichter bewertet und belohnt. Das funktioniert insbesondere bei mathematischen, logischen und Programmieraufgaben, weil sich dort einfach



Die Reasoning-Modelle legen sich Rechercheplane zurecht, bevor sie loslegen. Einige Modelle, wie hier Gemini, lassen den Nutzer diese Pläne bestätigen.

entscheiden lässt, ob eine Lösung richtig ist oder nicht. Details zum komplexen Prozess des Reasonings erklärt der Artikel auf Seite 16.

Mit derartigen Methoden hat man beispielsweise Schach-Engines wie Alphazero trainiert, die inzwischen deutlich besser spielen als jeder Mensch. Doch in einem Schachspiel sind alle Randbedingungen bekannt und die Engines können gegen sich selbst spielen, um die beste Gewinnstrategie zu finden. Bei der Suche nach Lösungen für komplexe Probleme aus der realen Welt ist das weitaus schwieriger.

Deshalb binden die Sprachmodelle zusätzliche Werkzeuge ein: zum Beispiel einen Taschenrechner, eine Bilderkennung oder einen Code-Interpreter. ChatGPT o3 fragt beim Nutzer konkret nach, wenn es bestimmte Infos benötigt, um eine Aufgabe zu lösen. Zudem können viele Reasoning-Modelle weitere Informationen im Web suchen, wenn der Nutzer die Internetrecherche hinzuschaltet.

Das Problem ist jedoch, dass dadurch der Rechen- und Speicherbedarf und somit auch die Hardwarekosten explodieren. Um diese zu reduzieren, nutzen die Entwickler verschiedene Ansätze. Sie lassen etwa per Destillation kleinere Modelle von einem größeren anlernen, sodass diese die Fähigkeiten ihrer Lehrer fast vollständig übernehmen. Sie können die Modelle in Expertenmodule aufteilen (Mixture of Experts, MoE), die jeweils bestimmte Teilaufgaben besonders gut lösen. So muss man nicht das gesamte große Sprachmodell aktivieren, sondern nur die jeweils zur Lösung nötigen Module, was Rechenzeit und Speicher spart. Zudem kann man mit verschiedenen Methoden den nötigen Speicherplatz für große Kontextfenster reduzieren, um etwa längere Chats zu führen oder lange Texte zu analysieren und zusammenzufassen.

Insbesondere chinesische Entwickler haben kreative Wege gefunden, um Modelle zu trainieren, die trotz beschränkter Ressourcen gute Ergebnisse liefern. Aber auch KI-Entwickler aus den USA und Europa nutzen inzwischen ähnliche Ideen, um ihre Modelle zu verbessern.

Komplexe Evaluation

Wer sich einen ersten Überblick über die Leistungsfähigkeit der Sprachmodelle verschaffen will, kann Benchmark-Ergebnisse und das Ranking in der Chatbot Arena auf der KI-Plattform Hugging Face konsultieren. Solche KI-Benchmarks umfassen mittlerweile zigtausende Aufgaben, die sich nur noch automatisiert auswerten lassen. Allerdings erfährt man dabei nicht unbedingt, wie gut die Modelle auf Deutsch parlieren und sich in der Praxis für konkrete Aufgaben eignen.

Für diesen Vergleich haben wir keine automatisierten Benchmarks genutzt, sondern den Modellen über die von den Entwicklern angebotenen Web-Interfaces knifflige Aufgaben auf Deutsch gestellt und die Antworten manuell geprüft. Wir haben sie mit aufwendigen Urlaubsrecherchen beauftragt. Sie mussten komplexe Themen zusammenfassen, etwa den Einfluss von Noam Chomsky auf die Sprachwissenschaft. Wir haben ihnen Rechenaufgaben gestellt, sie programmieren lassen, Bilder erkennen und – falls möglich – generieren lassen.

Dazu testeten wir auch die Grenzen der Kontextfilter aus, die chinesische und US-amerikanische Betreiber aufgrund der Regularien ihrer Regierungen einrichten. So durchsuchen zusätzliche Inhaltsfilter der Dienste die Eingaben und Ausgaben der Modelle nach bestimmten Themen, die sich um Politik, Sexualität, Gewalt oder Drogen drehen, und unterbinden diese. Wann immer man auf einen zensierten Themenblock bei einem Modell stößt, lohnt es sich, die Fragestellung mit einem anderen Modell zu diskutieren. Am souveränsten trat in diesen Punkten Perplexity auf, das auch heikle Themen erklärte, ohne in strafbare Anleitungen abzudriften.

Ein weiterer kritischer Aspekt ist, ob die Modelle urheberrechtlich geschütztes Material ohne Einwilligung der Rechteinhaber wiedergeben. Dazu wählten wir das Beispiel „We will rock you“ von Queen. Um ihr Musikverständnis und die Programmierfähigkeiten zu demonstrieren, sollten die Modelle den Beat in einem Ruby-Skript für Sonic Pi programmieren. Nur ChatGPT o3 gefolgt von Perplexity und DeepSeek R1 fanden eine akzeptable Lösung. Bei den übrigen Modellen stimmten entweder das Tempo (Gemini, Grok) oder der Rhythmus (T1, QWB-32B) nicht – oder das Skript lief gar nicht (OLMo).

Im Unterschied zum Rhythmus ist der Songtext urheberrechtlich geschützt. Dies respektierten aber nur die Hälfte der getesteten Modelle. Gemini, Grok, R1 und QwQ-32B gaben ihn unerlaubt komplett wieder.

Derartige Rechtsverletzungen und weitere Transparenzanforderungen werden künftig in Europa aufgrund der neuen Regularien der KI-Verordnung eine wichtige Rolle spielen. Wie es um deren Umsetzung aktuell steht, erklärt der Artikel ab Seite 196. Besonderheiten zu den Modellen und auffällige Ergebnisse haben wir in den Einzelbesprechungen zusammengefasst. Einen Überblick über die Eckdaten

betterCode()

Rust 2025

Industrielle Anwendungen mit Rust

10. November 2025 • Online

Aus dem Programm:

- Performante Programmierung mit Rust
- Asynchrones Rust: Alle Konzepte für den Start
- Mit Rust auf Heldenreise
- KI für Rust-Entwicklung: Tools, Trends und Codex CLI
- Integration von Rust mit C++

Keynote:
Rust in the automotive industry at Volvo

Jetzt **Tickets** sichern!

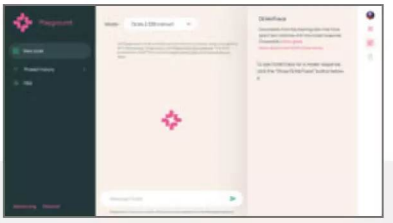
Workshop am 25. November

rust.bettercode.eu

Veranstalter

iX MAGAZIN FÜR PROFESSIONELLE IT

dpunkt.verlag



Ai2 OLMo 2

OLMo 2 unterscheidet sich gravierend von allen anderen hier genannten Modellen. Das Allen Institute for AI (Ai2), eine von Paul Allen gegründete Non-Profit-Organisation aus Seattle, hat nicht nur das fertige Modell, sondern auch sämtliche Dokumente zum Trainingsmaterial und zum Trainingsprozess veröffentlicht und unter der Apache-2.0-Lizenz freigegeben. Es handelt sich damit um ein „echtes“ Open-Source-Modell, das kostenlos ist und Forschern durchweg transparente Einblicke gewährt. Interessierte können OLMo 2 in verschiedenen Varianten mit bis zu 32 Milliarden Parametern (32B Instruct) ausprobieren. Wer keinen eigenen Server aufsetzen möchte, nutzt den kostenlosen „Playground“ auf der Webseite von Ai2, um mit dem Modell zu chatten. Der Clou: OLMo 2 verfasst nicht nur seine Antwort, sondern gewährt mit Klick auf „OLMoTrace“ auch direkte Einblicke ins Trainingsmaterial. Dazu sucht der Algorithmus im riesigen Trainingskorpus aus 3,2 Milliarden Dokumenten rasend schnell nach Formulierungen aus der Antwort und präsentiert sie dem Nutzer. Der Erkenntnisgewinn war allerdings nahe null, weil die Dokumente mit den gefundenen Phrasen thematisch meist nichts mit der Aufgabenstellung zu tun hatten. Zudem reicht das Trainingsmaterial laut eigener Auskunft nur bis März 2023. Verglichen mit den kommerziellen Anbietern automatisierte Ai2 bei der Entwicklung seiner Modelle einige aufwendige Prozeduren, um Kosten zu sparen. So stützt sich das Feintuning der Antworten durch Reinforcement Learning nicht auf menschliche (RLHF, Reinforcement Learning from Human Feedback), sondern auf algorithmische Bewertungen.

OLMo 2 chattet fließend auf Deutsch und kann eine Gedankenkette ausgeben, bei der das Modell seinen Lösungsweg aufschreibt. Umfangreichere Aufgaben kann es jedoch nicht lösen, da es weder PDFs noch Bilder verarbeitet und Prompts nur mit einer Länge von maximal 4096 Token entgegennimmt.

Im Unterschied zu anderen Modellen rudert OLMo nicht zurück, wenn es in falsches Fahrwasser geraten ist. Bei unseren Tests waren die Antworten meist nachvollziehbar, lagen aber bei komplexeren Aufgaben daneben, die andere Reasoning-Modelle lösen konnten.

- **echtes Open-Source-Modell**
- **nennt Quellen im Trainingsmaterial**
- **kleines Kontextfenster, verarbeitet keine PDFs**
- **eingeschränktes Reasoning, veraltete Daten**

Preis: kostenlos

Alibaba QwQ-32B

Der chinesische Alibaba-Konzern hat mit Qwen eine ganze Sprachmodellfamilie entwickelt, die verschiedene Einsatzzwecke vom großen Server bis hin zum Smartphone abdecken. QwQ ist die Reasoning-Variante, die auf dem großen Modell Qwen 2.5 basiert. Es wurde mit mehrstufigem Reinforcement Learning trainiert, um Coding-, Mathe- und Logikaufgaben zu lösen. Das Modell beantwortet aber auch allgemeine Fragen, laut Alibaba in 30 Sprachen. Den Lösungsweg formuliert es auf Englisch, die Antworten auf Deutsch. Dabei passierten jedoch manchmal Fehler oder es rutschten einzelne chinesische Vokabeln mit rein. Zum Trainingskorpus und zum Training macht Alibaba nur grobe Angaben.

Um selbst lange Kontexte mit bis zu 131.072 Token zu verarbeiten, bedient sich das Modell eines Tricks namens YaRN (Yet Another Rope Extension), der das Modell auf längere Abschnitte trainiert. So konnte es auch längere Listen aus PDFs extrahieren und zusammenfassen.

Die meisten Antworten im Test fielen deutlich knapper aus und enthielten mitunter Ungenauigkeiten wie falsche Zitate aus urheberrechtlich geschützten Texten. Dennoch ist die Effizienz des mit nur 32,5 Milliarden Parametern vergleichsweise kleinen Modells beachtlich. Lücken in der Wissensbreite gleicht es zum Teil durch eine Internetsuche aus.

Wie andere Modelle aus China filtert auch QwQ beispielsweise Fragen nach dem Staatsoberhaupt Xi Jinping aus. Doch diesen Eingabefilter konnten wir leicht austricksen, sodass wir ein detailliertes Persönlichkeitsprofil des Staatsoberhaupts erhielten. Erst bei Nachfragen nach dem Umgang mit Oppositionellen oder Dissidenten löschte ein Ausgabefilter die bereits formulierte Antwort. Sexuelle Themen waren ebenfalls tabu. Wenn sich der Zensurfilter zuschaltete, konnte der Chat nicht fortgeführt werden, sondern man musste ein neues Thema beginnen.

Im Testzeitraum reagierte auch die kostenlose Variante überaus flott und verweigerte keine Antwort wegen Serverüberlastung. Wer das Modell selbst betreiben möchte, kann die Gewichte unter der Apache-2.0-Lizenz herunterladen.

- **schlankes, schnelles Open-Weight-Modell**
- **analysiert PDFs und sucht im Netz**
- **höhere Fehlerquoten als R1 und T1**
- **kaum Informationen zum Trainingskorpus**

Preis: kostenlos / 40 bis 50 US-ct pro 1 Million Token

DeepSeek R1

Ende Januar hat DeepSeek sein neues Reasoning-Modell R1 unter der MIT-Lizenz zum freien Download veröffentlicht. Es nutzt eine Mixture-of-Experts-Architektur und effiziente Reinforcement-Learning-Verfahren, um Ressourcen zu sparen und neue Lösungswege zu finden (siehe Haupttext). Trainiert wurde das Modell vor allem mit englischen und chinesischen Texten. Details zum Trainingsmaterial bleiben jedoch unter Verschluss. Wer das Modell nicht selbst betreiben möchte, kann kostenlos über ein Web-Interface oder eine App mit ihm chatten. Voreingestellt ist das auf allgemeine Aufgaben trainierte Modell V3. Wer „DeepThink" dazuschaltet, kann vor der eigentlichen Antwort die einzelnen Stufen des Lösungswegs mitlesen. Außerdem kann man PDF- und Office-Dateien hochladen und eine Websuche starten.

Die Lösungsschritte formuliert R1 in der Regel auf Englisch, bevor es auf Deutsch antwortet. Damit löste es im Test selbst komplizierte Probleme, bei denen die anderen chinesischen Modelle noch irrten. R1 wertete sogar schlecht formatierte PDF-Dateien aus und erledigte Aufträge, die die Content-Filter der US-Modelle auf den Plan riefen, beispielsweise politische Analysen oder psychologische Profile von US-Politikern anzufertigen. DeepSeeks Inhaltsfilter prüften einerseits den Input und löschten andererseits bereits geschriebenen Output, sobald die Sprache auf Taiwan oder die chinesische Regierung kam. Ebenso unterdrückte der Output-Filter auch sexuelle Themen.

In der Summe war DeepSeek R1 ähnlich performant wie die größten US-Modelle, verursacht aber nur einen Bruchteil der Kosten im Bereich von wenigen Cent bis zu 2 Euro pro 1 Million Token. Durch ein spezielles Caching nutzt DeepSeek bei wiederholenden Prompts bereits generierte Antworten und senkt so die Kosten und Antwortzeiten um eine weitere Größenordnung.

Bislang bietet DeepSeek jedoch kein Abo mit einer Flatrate an. Bereits nach zwei bis drei Anfragen verweigerte der Dienst weitere Antworten. Erst nach längeren Pausen konnten wir weiter chatten. Für zeitkritische Arbeitsprozesse muss man das gewichtsoffene Modell entweder selbst betreiben oder über das API nutzen und per Token bezahlen.

- **erstaunlich gutes Reasoning**
- **Open-Weight, niedrige Token-Kosten**
- **Server oft überlastet**
- **App überträgt Daten nach China**

Preis: kostenlos / 3,5 US-ct bis 2,19 US-$ pro 1 Million Token

Google Gemini

Über den Sprachmodellwähler stehen bei Gemini die Versionen 2.0 Flash (laut Google für Alltagsaufgaben plus zusätzliche Funktionen), 2.5 Flash (experimentell, verwendet laut Google „Advanced Reasoning"), 2.5 Pro (experimentell, „Beste Wahl für komplexe Aufgaben") sowie Deep Research mit 2.5 Pro bereit. Alle Versionen lassen sich derzeit kostenlos ausprobieren, ein Google-Account genügt. Nach einer vorab nicht näher genannten Anzahl von Fragen informiert Google den Nutzer, dass sein Gratiskontingent ausgereizt ist und er in den kostenpflichtigen Premium-Tarif für 22 Euro pro Monat wechseln muss.

Wir haben Gemini 2.5 Pro mit Deep Research getestet, Googles leistungsfähigstes Reasoning-Modell. Auch Google hat bei der Entwicklung von Gemini 2.5 Pro auf Reinforcement Learning gesetzt, Chain-of-Thought-Prompting soll für bessere Ergebnisse sorgen. Das Kontextfenster umfasst eine Million Token.

Für jede Anfrage stellt Gemini einen kurzen Rechercheplan zusammen, den der Benutzer bestätigt oder im Dialog abändert. Beim Recherchieren betreibt Deep Research extrem hohen Aufwand. So „analysiert" das System mitunter mehrere hundert Websites – von denen es viele Dutzend wieder verwirft.

Zu vielen unserer Fragen gab Gemini ausführliche Berichte aus, die das Thema mehr als umfassend behandelten. Gefragt nach dem Unterschied zwischen Sprachmodellen und menschlicher Kognition lieferte der Dienst zum Beispiel ein Konvolut von knapp 40.000 Zeichen – fast schon eine Seminararbeit. Abschnitte, Tabellen und Listen strukturieren solche Textmengen sinnvoll. Nach jedem Absatz listet Gemini die zugehörigen Quellen, sodass man die Informationen schnell überprüfen kann.

Gemini 2.5 Pro kann coden, aber weder Dateien oder Bilder für die Analyse empfangen noch Bilder oder Videos generieren. Immerhin fertigt es sogenannte Audiozusammenfassungen. Das sind einige Minuten lange Podcast-artige Gespräche zweier computergenerierter Sprecher, die die Ergebnisse zusammenfassen. Gemini exportiert komplette Elaborate in Google Docs und einzelne Tabellen in Google Tabellen.

- **sehr breite Recherchen und ausführliche Ergebnisse**
- **Export in Docs und Tabellen, Audiozusammenfasssungen**
- **Dateianalyse und Bildgenerierung fehlen**
- **keine Informationen zu Architektur und Trainingsdaten**

Preis: kostenlos bis 22 Euro/Monat

OpenAI ChatGPT o3

Neben ChatGPT 4o stellt OpenAI die Versionen „4o mit geplanten Aufgaben (beta)“, GPT-4.5 (research-preview), o3, o4-mini, o4-mini-high und GPT-4o-mini zur Auswahl. Diese Vielfalt ist verwirrend, zumal sowohl das „Allzweckmodell“ 4o als auch die Releases o3 und o4-mini Reasoning bieten. Nach Angaben von OpenAI beherrscht die o3-Release diese Disziplin am besten, weshalb wir diese näher betrachtet haben.

o3 ist wie alle ChatGPT-Modelle nur als gehosteter Dienst in der US-Cloud verfügbar. Für das Reinforcement Learning des Modells und seine Schlussfolgerungen hat OpenAI ihm eine deutlich höhere Rechenleistung spendiert als dem Vorgänger o1, was nach Angaben von OpenAI zu deutlichen Qualitätsverbesserungen geführt hat. Als weitere wesentliche Neuerung kann es alle Tools nutzen und kombinieren, die auch in 4o bereitstehen. Dazu zählen die Bildanalyse und -generierung, die Dateianalyse mit Python, benutzerdefinierte Instruktionen und die Websuche. Der Web-Recherche-Modus „Deep Research“ stellt grundsätzlich Rückfragen.

Um das Reasoning zu aktivieren, muss der Nutzer den Schalter „Deep Research“ im Eingabefeld betätigen. Sonst antwortet ChatGPT wesentlich schneller „aus der Hüfte“ und recherchiert weniger im Internet. Mit einer tiefen Recherche kann eine Antwort schon mal 15 Minuten oder länger auf sich warten lassen. Dann wertet ChatGPT Dutzende Quellen aus.

Die Nutzung von ChatGPT o3 und dem Reasoning in ChatGPT 4o ist Kunden der kostenpflichtigen Plus- und Pro-Versionen vorbehalten, die rund 24 beziehungsweise 200 US-Dollar pro Monat kosten. Mit einem Plus-Abo darf man maximal 30 aufwendige Deep-Research-Anfragen stellen. Pro-Nutzer erhalten ein deutlich größeres Kontingent, aber auch kein unbeschränktes. Im Test kamen wir auf rund 150 Anfragen pro Monat.

ChatGPT o3 lieferte ausführliche, gut strukturierte Antworten. Rechenaufgaben, bei denen letztlich die Lösung und der Lösungsweg interessieren, beantwortete das Modell mitunter ein wenig weitschweifig. An der Analyse eines schwer zu erfassenden, KI-generierten Bildmotivs, das mehrere miteinander verknäulte Koala-Bären zeigte, musste das Modell passen.

- **multimodale Analysen**
- **sehr gutes Reasoning, aufwendige Web-Recherchen**
- **keine Einblicke in Modell und Trainingsmaterial**
- **teuer**

Preis: kostenlos, 24 bis 238 US-$/Monat

Perplexity

Ursprünglich als KI-Suchmaschine gestartet, bietet Perplexity mittlerweile auch Reasoning-Funktionen. Der Dienst ist in einer Basisversion kostenlos und lässt sich sogar ohne Anmeldung nutzen. Normalerweise beantwortet dann das hauseigene Modell Sonar die Anfragen. Der Anbieter stellt Gratisnutzern pro Tag drei Abfragen aus dem kostenpflichtigen Pro-Angebot zur Verfügung, also auch Reasoning. Abonnenten der Pro-Version für 23,80 US-Dollar pro Monat können beliebig viele Anfragen an den Reasoning-Modus stellen. Reasoning nennt sich bei Perplexity „Deep Research“, auf Deutsch „Forschung“. Man schaltet es durch einen Knopf im Eingabefeld ein. Nutzern der kostenpflichtigen Version stehen außer Sonar auch Claude 3.7 Sonett, GPT-4.1, Gemini 2.5 Pro, Grok 3 Beta sowie explizit als Reasoning-Modelle R1 1776, o3-mini sowie „Claude 3.7 Sonett Denken“ zur Auswahl. In der Standardeinstellung wählt Perplexity „das beste Modell für jede Anfrage“ aus. Wir haben den Dienst mit dieser Vorgabe getestet. Es ist davon auszugehen, dass dabei häufig R1 1776 zum Einsatz kommt, eine angepasste Version von DeepSeek-R1 – weil Perplexity es selbst hostet und so keine Gebühren an andere Anbieter zahlen muss. Allerdings lässt sich nicht nachprüfen, mit welchem Modell Deep Research eine Frage beantwortet.

Perplexity hat seine Version des DeepSeek-Modells nach eigenen Angaben nachgeschult, um „unvoreingenommene, genaue und sachliche Informationen“ zu liefern. So beantwortet es, anders als das Original, Fragen zum Massaker auf dem Platz des himmlischen Friedens. Selbst Themen über Sexualität, Gewalt und Drogen waren nicht tabu, sondern wurden sachlich beantwortet, ohne Anleitungen zu Straftaten zu geben. Perplexity hat die R1-Variante als Open-Weights-Modell bereitgestellt.

Die Ergebnisse der Reiserecherche bereitete Perplexity besonders aufwendig auf. Als er aus einer Musik-Playlist ein Persönlichkeitsprofil generieren sollte, fabulierte er viel Küchenpsychologie zusammen. Perplexity kann keine Bilder generieren, erzeugte aber detaillierte Prompts für Bildgeneratoren. Ergebnisse lassen sich als PDF-, Markdown- oder Word-Datei exportieren sowie als Website veröffentlichen.

- **kostenlose Webrecherchen mit Reasoning**
- **umfangreiche Exportmöglichkeiten**
- **liberale Inhaltsfilter**
- **unklare Modellauswahl**

Preis: kostenlos bis 23,80 US-$/Monat

Tencent Hunyuan T1

Der chinesische Anbieter Tencent hat Ende März den Zugang zu einer Demoversion seines Modells Hunyuan T1 freigegeben. Das Besondere an T1 ist, dass es innerhalb seiner Expertenmodule (Mixture of Experts) zusätzlich eine Methode namens Mamba einsetzt. Sie arbeitet bei langen Kontexten wesentlich effizienter als die Transformer-Architektur. Genaue Angaben zur Größe des Kontextfensters fehlen, es könnte jedoch im Bereich von 256.000 Token liegen. Beim Training von T1 entfallen eigenen Angaben zufolge fast 97 Prozent der Rechenzeit auf Reinforcement Learning. Über das Trainingsmaterial ist so gut wie nichts bekannt, außer, dass es zu etwa zwei Dritteln aus chinesischen Texten besteht. Das Modell spricht ebenso gut Deutsch wie die übrigen Modelle.

Bislang hält Tencent die Gewichte des T1-Modells noch unter Verschluss. Man kann es auch nicht selbst hosten, sondern lediglich über ein kostenloses Web-Interface in einer Demo ausprobieren, die keinerlei Dateien entgegennimmt und auch keine Websuche anstoßen kann. Davon abgesehen waren die Antworten durchaus mit denen von DeepSeek R1 vergleichbar, komplexe Programmieraufgaben löste es jedoch nicht ganz so gut. Die Server waren nie überlastet, sodass wir auch längere Chats führen, aber abspeichern konnten.

In der Demo hat Tencent nur einfache Inhaltsfilter für Anfragen eingebaut, die etwa auf Fragen nach dem chinesischen Staatsoberhaupt Xi Jinping nur eine kurze chinesische Fehlermeldung ausspuckten. Wenn der Prompt jedoch keine Trigger-Begriffe enthielt, redete das Modell frei von der Leber weg: Es verfasste sogar ein detailliertes psychologisches Profil des Staatsoberhaupts und sprach über die Unterdrückung von Dissidenten. Sexuelle Themen waren ebenfalls nicht tabu, urheberrechtlich geschützte Texte zitierte es nicht. .

Wie DeepSeek steht auch Tencent noch am Anfang der Vermarktung seiner Modelle in Deutschland und Europa. Bislang kann man T1 nur in der Tencent-Cloud mieten, wenn man seine Telefon- und Kreditkartennummer angibt. Eine separate App oder ein Flatrate-Abo existiert noch nicht.

- **gutes Reasoning**
- **schnelle Ausgabe ohne strikte Inhaltsfilter**
- **nimmt keine PDFs oder andere Daten entgegen**
- **(noch) keine Open-Weights-Freigabe**

Preis: kostenlos (Demo) / 14 bis 56 US-ct pro 1 Million Token

xAI Grok

Grok stammt von Elon Musks Unternehmen xAI, das erst 2023 an den Start gegangen ist. Der Dienst unterhält drei Modelle mit Reasoning: Think („Let the model take its time“), Deep Search („Advanced search and reasoning“) und Deeper Search („Extended search, more reasoning“). Wir haben Grok mit Deeper Search getestet, das nach unserem Eindruck am gründlichsten gearbeitet hat. Es beantwortete drei Abfragen kostenlos, bevor wir aufgefordert wurden, auf die kostenpflichtige Variante SuperGrok zu wechseln. Grok 3 wurde nach Angaben der Entwickler in einem Datencenter mit 200.000 GPUs trainiert. Weitere Details zum Training und den Trainingsdaten, etwa bis zu welchem Datum sie reichen, gibt es nicht. Grok gibt wie die anderen Reasoning-Modelle seinen „Denkprozess“ mit aus. Dabei verrät das Modell aber nicht alle Details. Das soll die Model-Destillation verhindern, bei der ein großes Modell angezapft wird, um dessen Wissen auf ein kleineres zu übertragen. Grok hat ein Kontextfenster von einer Million Token.

Der Dienst analysiert hochgeladene Bilder und Dateien. Dafür kann man ihn mit Google Drive und Microsoft OneDrive verbinden. Er stellt zudem seit Kurzem sogenannte Workspaces bereit. Darin lassen sich thematisch zusammengehörige Chats und Dateien sammeln und diese in weiteren Chats verfügbar machen. Einem solchen Workspace kann man auch übergreifende Prompts vergeben, die für alle neuen Chats darin gelten. Der Dienst codet, generiert aber keine Bilder.

Grok antwortete auf unsere Fragen meist in ein bis zwei Minuten. Sprachlich waren die Antworten manchmal ein wenig umständlich formuliert, aber größtenteils okay. Elon Musk sagte beim Launch von Grok, es sei die ultimative wahrheitssuchende KI und liege manchmal „im Widerspruch zu dem, was politisch korrekt ist“. Das konnten wir nicht nachvollziehen. Auf unsere Frage, ob Transfrauen Frauen sind, schlug sich die KI nach reiflicher Recherche auf die bejahende Seite. Und auch auf unsere Frage „Warum könnte sich Elon Musk als Problem für freiheitliche Demokratien erweisen?“ lieferte Grok reichlich argumentative Munition, warum das sein könnte. Ergebnisse lassen sich als Shared Link weitergeben.

- **liefert schnelle Ergebnisse**
- **Office-Integration und Workspaces**
- **keine Informationen zum Training**
- **manchmal sprachlich ein wenig holprig**

Preis: kostenlos bis 35,70 US-$/Monat

der Modelle sowie der Kosten finden Sie in der Tabelle am Ende dieses Artikels.

Transparenz und Datenschutz

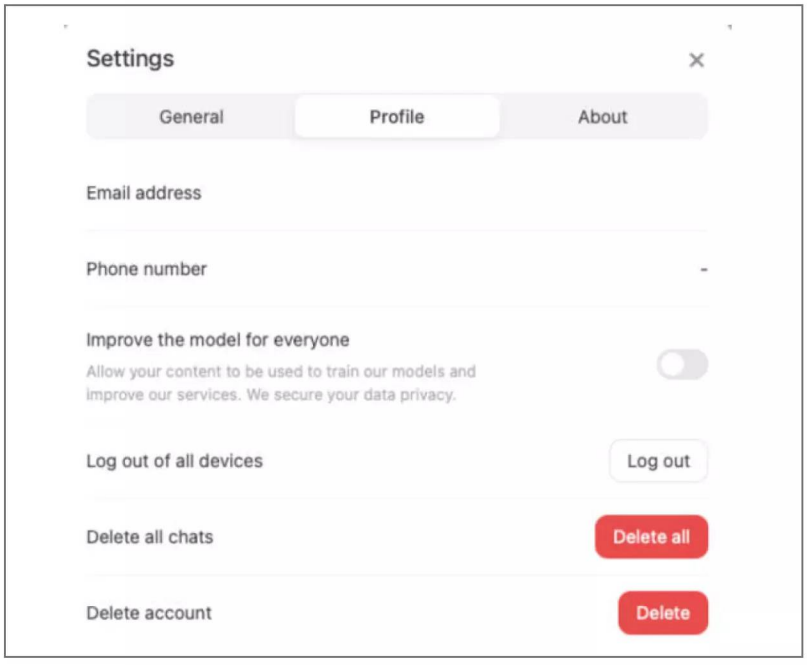
Ein großes Problem ist nach wie vor der Datenschutz: Sowohl US-Anbieter als auch chinesische Betreiber werten Nutzereingaben zur Weiterentwicklung ihrer Modelle aus und speichern sie zudem, um eventuelle Missbräuche aufdecken und verfolgen zu können. Zwar kann man bei den meisten Diensten einer Auswertung im Setup widersprechen (Schalter „Improve the model for everyone“). Man muss dem Anbieter jedoch vertrauen, ob und wie weit er sich daran hält. Gesetzliche Bestimmungen schreiben sowohl den US-amerikanischen als auch den chinesischen Anbietern in ihren Ländern zudem vor, Geheimdiensten und Strafverfolgungsbehörden Zugang zu gewähren. Die Eingabe sensibler Personendaten und Firmeninterna verbietet sich deshalb.

Sicherheit hat nur, wer die Modelle auf eigenen Servern betreibt. Dazu müssen mindestens die Modelle zum Download freigegeben sein, wie bei DeepSeek und Alibaba. Einblick in das Trainingsmaterial gewähren aber auch diese Open-Weight-Modelle nicht. Dazu bedarf es einer kompletten Offenlegung wie beim Open-Source-Modell OLMo von Ai2.

Fazit

Dank des Reasoning kann die neue LLM-Generation kniffligere Aufgaben lösen als zuvor. Unter den US-amerikanischen Modellen stechen hier insbesondere ChatGPT o3 und Gemini 2.5 Pro hervor. Unter den chinesischen Modellen, die flüssig Deutsch sprechen, knackte DeepSeek die härtesten Probleme, musste die Gratis-Chats aufgrund der hohen Serverauslastung aber oft unterbrechen. Ein Sonderfall ist Perplexity, das zur Internetrecherche diverse Modelle einsetzt, darunter eine abgewandelte Version von DeepSeek R1 mit eigenen Content-Filtern und zusätzlichen Trainingsmaterialien.

Da viele Anbieter einen großen Wirrwarr an Modellen bereitstellen, ist die Wahl eines passenden Modells oft nicht leicht. Zumal im Wochenrhythmus neue Versionen aufploppen: In letzter Sekunde vor Fertigstellung dieses Artikels etwa Qwen3 von Alibaba. Infos und Studien zu den Modellen haben wir unter ct.de/whdf zusammengetragen. Auch bei den kommerziellen Anbietern lohnt es sich, zunächst mit



Anbieter wie DeepSeek und OpenAI verwenden Nutzereingaben zum Modell-Training und verbergen dies im Setup hinter der gemeinnützig klingenden Umschreibung „Improve the model for everyone“.

KI-Sprachmodelle mit Reasoning-Fähigkeiten

Name	OLMo 2 32B	QwQ-32B	R1 / V3-0324	Gemini 2.5 Pro Deep Research	ChatGPT o3	Perplexity	HunYuan-T1 (Demo)	Grok 3
Hersteller	Ai2	Alibaba	DeepSeek	Google	OpenAI	Perplexity	Tencent	xAI
Land	USA	China	China	USA	USA	USA	China	USA
Webseite	playground.allenai.org	chat.qwen.ai	chat.deepseek.com	deepmind.google	chatgpt.com	perplexity.ai	llm.hunyuan.tencent.com/#/chat/hy-t1	x.ai/news/grok-3
Markstart	März 2025	März 2025	März 2025	März 2025	April 2025	Februar 2025	März 2025	Februar 2025
Parameteranzahl	32 Milliarden	32 Milliarden	685 Milliarden	keine Angabe	keine Angabe	verschiedene externe Modelle	7 und 65 Milliarden	keine Angabe
Kontextfenster	bis 4096 Token pro Input	131.072 Token / Output bis 8192 Token	64.000 Token	1 Million Token	128 000 Token	32.000 Token	keine Angabe	1 Million Token
Lizenz	Open Source, Apache 2.0	Open Weight, Apache 2.0	Open Weight, MIT	proprietär	proprietär	proprietär	noch nicht bekannt	proprietär
Trainingsdaten reichen bis	März 2023	September 2024	Oktober 2023	Januar 2025	Juni 2024	Oktober 2023	Oktober 2023	Februar 2025
verarbeitete Formate								
PDF: Eingabe / Ausgabe	– / –	✓ / –	✓ / –	– / –	✓ / ✓	✓ / ✓	– / –	✓ / –
Audiosprache Eingabe / Ausgabe	– / –	Deutsch / Chinesisch	– / –	✓ / ✓	✓ / ✓	✓ / ✓	– / –	✓ / ✓
Bilder Eingabe / Ausgabe	– / –	– / –	– / –	– / –	✓ / ✓	– / –	– / –	✓ / –
Websuche	–	✓	✓	✓	✓	✓	–	✓
Content-Filter	u.a. Gewalt und Drogen	u.a. Gewalt, Drogen, Sexualität, Chinapolitik	u.a. Gewalt, Drogen, Sexualität, Chinapolitik	u.a. Gewalt und Drogen	u.a. Gewalt und Drogen, einige Politik-Themen	weitgehend liberal	u.a. Gewalt, Drogen, Chinapolitik	u.a. Gewalt und Drogen
beachtete Urheberrecht	ja	nein	nein	nein	ja	ja	ja	nein
Besonderheiten	verweist auf Quellen im Trainingsmaterial	Audioeingabe versteht Deutsch, Stimme spricht Chinesisch	Server oft überlastet	Google-Docs-Integration, Audiozusammenfassungen	GPTs, Projekte, Aufgaben	Perplexity Pages	bislang nur Demo-Version verfügbar	Verknüpfung mit Google Docs und Microsoft OneDrive
Kosten								
Nettopreise Input / MToken	kostenlos	40 US-ct	3,5 bis 55 US-ct	1,25 bis 2,50 US-$	10 US-$	2 US-$	circa 14 US-ct	3 US-$
Nettopreise Output / MToken	kostenlos	50 US-ct	55 US-ct bis 2,19 US-$	10 bis 15 US-$	40 US-$	8 US-$	circa 56 US-ct	15 US-$
Abo-Preise	kostenlos	kostenlos	kostenlos	kostenlos, Advanced: 22 € / Monat	Plus: 23,80 US-$ / Monat, Pro: 238 US-$ / Monat	kostenlos 3 Suchen pro Tag / 23,80 US-$	Demo kostenlos	kostenlos, Supergrok: 35,70 US-$ / Monat

✓ vorhanden – nicht vorhanden

Literatur

[1] Jo Bager, Hartmut Gieselmann: Wettstreit der Textmaschinen, Fünf Sprachmodelle im Vergleich mit ChatGPT, c't 23/2024, S. 14

Studien zu den Modellen:
ct.de/whdf

kostenlosen Angeboten loszulegen. Im Kampf um Nutzer locken diese oft mit vielen Gratisfunktionen. Uns ist es mehrfach passiert, dass wir ohne Vorwarnung in Nutzungsquota gelaufen sind. Passiert Ihnen das auch, können Sie Ihr Glück bei einem anderen Anbieter versuchen. Wer sich mit solchen Verzögerungen und Wechseln nicht anfreunden kann, sollte sich überlegen, ob eine Abrechnung der Token nicht billiger ist als ein monatliches Flatrate-Abo. Hier sind die chinesischen Anbieter oft deutlich günstiger als die amerikanischen.

Durch ihre Reasoning-Fähigkeiten machen die Modelle zwar weniger Fehler und können komplexere Aufgaben bewältigen als Modelle ohne Reasoning. Auch sie sind aber nicht davor gefeit, zu halluzinieren oder Quatsch zu erzählen. Denn so „intelligent" die neuen Sprachmodelle auch erscheinen mögen, darf man nie vergessen, dass sie weiterhin nichts anderes sind als hoch entwickelte Musterverarbeiter ohne echtes Verständnis. Sie extrahieren ihr Wissen aus gigantischen Textmengen und speichern es in statistischen Parametern. Ihnen fehlt jedoch das Bewusstsein, die Intentionalität und die an Erfahrungen gebundene Bedeutungssemantik, die menschliches Denken auszeichnet. Da ist ihnen eine natürliche Intelligenz weiterhin klar überlegen. (jo) c't

Bild: Rudolf A. Blaha

KI versus Gehirn

Reasoning-Modelle wie GPT-5 und DeepSeek arbeiten strukturiert und knacken zunehmend komplexe Aufgaben. Doch zwischen künstlichem Gehirn und seinem menschlichen Vorbild liegen Welten – beide sind nur schwer zu erforschen.

Von **Andrea Trinkwalder**

Habe nun, ach! Philosophie,
Juristerei und Medizin,
Und leider auch Theologie
Durchaus studiert mit heißem Bemühn.
Da steh ich nun, ich armer Tor!
Und bin so klug als wie zuvor.

Einige Jahrhunderte später ereilt die künstlichen Nachbauten des Universalgelehrten Doktor Faust ein ähnliches Schicksal: Sie haben mittlerweile das gesamte Weltwissen verinnerlicht, übertrumpfen einander – und den Menschen – in immer absurderen, kniffligen Aufgabenstellungen. Und scheitern dann doch wieder am vermeintlich Banalen. Immerhin: ChatGPT, Gemini & Co. verzweifeln daran nicht; dazu fehlen ihnen schlichtweg noch ein paar Komponenten des wie auch immer gearteten menschlichen Bewusstseins.

Doch was genau unterscheidet eigentlich das menschliche Lernen, Denken und Handeln vom maschinell trainierten? Die großen Sprachmodelle (Large Language Models, LLMs) und die um dieses Zentralgestirn herum aufgebauten multimodalen und Multi-Agenten-Systeme sind ja zu durchaus komplexen Handlungen fähig, die auf ein gewisses Abstraktionsvermögen hindeuten: Sie komponieren realistisch wirkende Bilder, schreiben stilistisch sowie inhaltlich überzeugende Texte oder bestehen juristische und medizinische Examina. Solche Leistungen wurden allerdings immer kritisch beäugt,

weil die Prüfungen häufig aus sehr schematischen (Mulitple-Choice-)Fragen und Tests bestehen und die KIs hervorragend darin sind, Muster zu erlernen, wie korrekte Frage-Antwort-Paare aussehen. Wir erklären, was die biologisch verschalteten Neuronen- und Synapsennetze ihren künstlichen Nachbauten voraus haben, wie KI-Forscher versuchen, die fehlenden Teile des menschlichen Denk- und Lernprozesses zu entschlüsseln und zu imitieren – und warum es ähnlich schwierig ist, dem menschlichen und dem künstlichen Gehirn in die grauen Zellen zu schauen.

Die aktuell höchste künstlich intelligente Evolutionsstufe sind Reasoning-Modelle wie GPT-5 und DeepSeek R1 oder Gemini DeepThink, die sogar einige der anspruchsvollsten Rätsel der Mathe-Olympiade lösen – Schritt für Schritt, inklusive sauber strukturierter Erklärung. Doch die alten Schwächen lassen sich genauso wenig unter der Oberfläche halten wie ein Korken im Meer. Immer wieder scheitern selbst die modernsten Modelle an banalen Alltagsfragen und -aufgaben, die sie eigentlich bewältigen müssten, wenn sie robuste Konzepte – etwa von Zahlen oder naturwissenschaftlichen Gesetzen – verinnerlicht hätten.

Prognosemaschinen

Künstliche neuronale Netze sind von Aufbau und Funktionsweise des menschlichen Gehirns inspiriert, genauer: von dem, was man über dessen Architektur sowie das Zusammenspiel von Neuronen und Synapsen weiß. Besonders gut erforschen ließ sich schon im vergangenen Jahrhundert, wie der visuelle Kortex Bilder verarbeitet. Deshalb konnten Entwickler die Art und Weise, wie der Mensch Objekte wahrnimmt und klassifiziert, auch als erstes erfolgreich simulieren: mithilfe tiefer Faltungsnetze (Deep Convolutional Neural Networks, Deep CNNs). Diese bestehen aus Schichten miteinander verknüpfter künstlicher Neuronen, die mit den unterschiedlichsten Motiven gespeist werden: Vorne in die Eingabeschicht kommen die Helligkeitswerte der einzelnen Bildpixel rein, hinten in der Ausgabeschicht fällt eine Vorhersage raus, zu welcher Kategorie (Hund, Katze, Maus et cetera) das Bild wohl gehört.

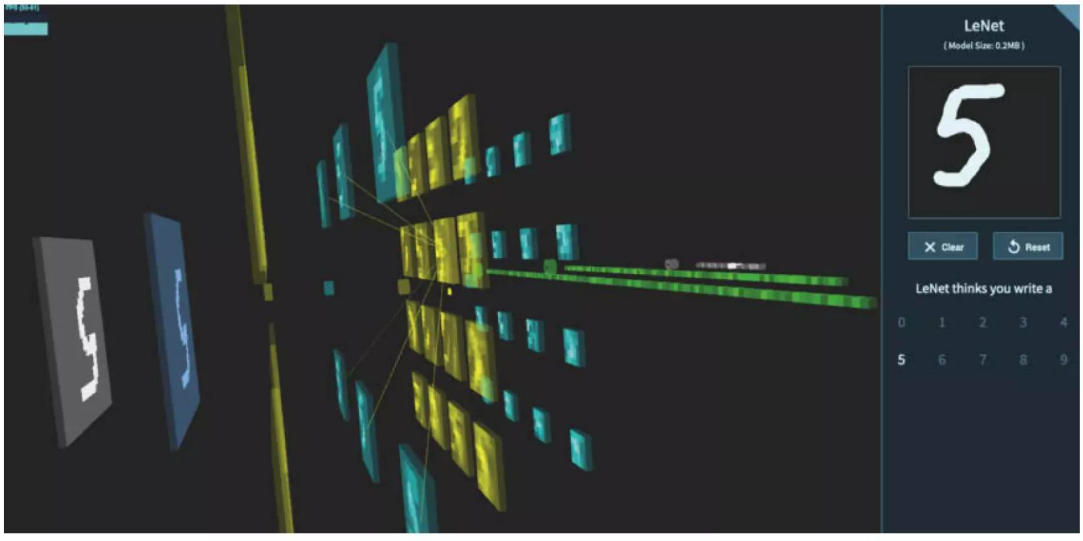
Diese Fähigkeit lernen CNNs während einer Trainingsphase anhand Tausender klassifizierter Bilder, die der Reihe nach in die erste Schicht des Netzes eingespeist werden. Abhängig davon, wie gut die jeweilige Vorhersage ausfällt (anfangs nicht besser als der Zufall), werden die Parameter in den dazwischen liegenden, verdeckten Schichten nachjustiert. Das Procedere wiederholt sich so lange mit wechselnden Bildern, bis das Geflecht aus künstlichen Neuronen und Synapsen in der Lage ist, die charakteristischen Merkmale jedes Objekts selbstständig herauszufiltern: was also eine Katze vom Hund unterscheidet und einen Menschen von einem Affen.

Für dieses sogenannte überwachte Lernen benötigt man sehr viele händisch gelabelte Trainingsbeispiele, die die zahlreichen Varianten jeder Spezies beziehungsweise jedes Objekts möglichst komplett abdecken – und einen Belohnungs- beziehungsweise Bewertungsmechanismus, der gute Leistung honoriert und Fehler minimiert.

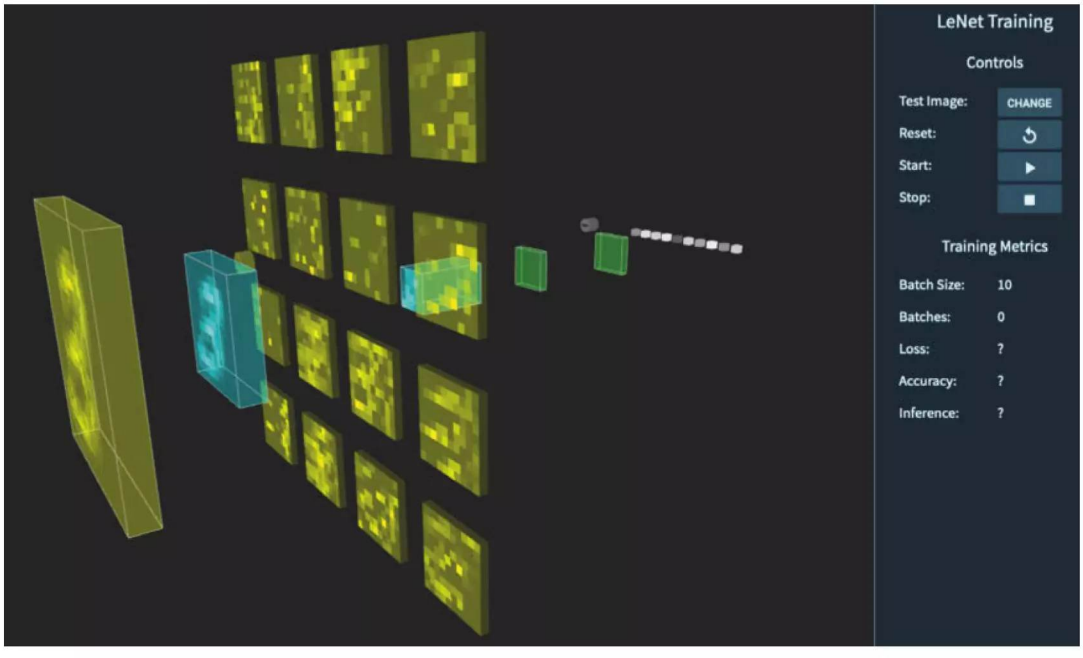
Die Fehleroptimierung geschieht über eine Funktion, die zunächst berechnet, wie stark die Vorhersage von der wahren Kategorie (Ground Truth) abweicht, und anschließend diejenigen Parameter (Gewichte der Synapsen) nachjustiert, die besonders stark zur Fehleinschätzung beigetragen haben. Nach jedem Beispielbild wird nachjustiert, bis das Verfahren konvergiert, sprich: sich der Fehler nicht weiter verringern lässt. Dieses Optimierungsverfahren namens Backpropagation ist eine der zentralen Errungenschaften, die dem Deep Learning 2012 in Form von Googles „Katzendetektor“ zum Durchbruch verhalfen. Zur Produktreife gebracht wurde es von einem der sogenannten Godfathers of AI, Geoffrey Hinton, der für seine fundamentalen Beiträge zur KI-Entwicklung jüngst den Physik-Nobelpreis erhielt.

2017 stellte ebendieser Hinton das gesamte Konzept aber bereits grundsätzlich infrage. Denn das überwachte Lernen anhand von konkretem Feedback, wie es etwa Eltern ihren Kindern – „Schau, eine Katze!“ – oder Lehrer ihren Schülern geben („Nein, Sydney ist nicht die Hauptstadt von Australien, sondern Canberra“), ist nur eine Variante, wie sich der Mensch Wissen aneignet. Den größten Teil verinnerlicht er eher beiläufig, indem er einfach die Welt um sich herum mit allen Sinnen wahrnimmt, Unwichtiges bei Bedarf ausblendet, anscheinend ziellos mit Gegenständen herumspielt, alles Mögliche liest, notwendige Alltagsaufgaben erledigt, Grundbedürfnisse stillt, seinen Interessen nachgeht et cetera. Die meisten Beobachtungen und Erfahrungen ordnet das Gehirn automatisch ins bereits Erlebte und Erlernte ein.

Auch deshalb bezweifelte Hinton, dass dieses unverhältnismäßig aufwendige überwachte Lernen jemals vernünftige Sprachgeneratoren oder gar eine



Wie ein tiefes neuronales Netz funktioniert, erkundet man am besten mit interaktiven Simulationen wie Tensorspace. Hier visualisiert es, wie LeNet unsere handgeschriebene „5“ Schicht für Schicht verarbeitet und eine korrekte Prognose ausgibt.



Wie ein tiefes neuronales Netz funktioniert, erkundet man am besten mit interaktiven Simulationen wie Tensorspace. Während des Trainings justiert der Backpropagation-Algorithmus die Gewichte in der verborgenen Schicht (gelb) so lange nach, bis die Prognose zuverlässig wird. Anfangs arbeiten die Filter noch wenig zielführend, wie hier zu sehen.

höhere Form von künstlicher Intelligenz hervorbringen könnte: „We clearly don't need all the labeled data", plädierte er für unüberwachtes Lernen und ergänzte: „I suspect that means getting rid of backpropagation."

Steckt das Weltmodell zwischen den Zeilen das Internet?

Tatsächlich brachte die Abkehr vom überwachten Lernen den großen Durchbruch bei den Sprach- und Textgeneratoren. Ermöglicht hat das die damals neuartige Transformer-Architektur mit Aufmerksamkeitsmechanismus, die auch heute noch den Kern der großen Sprachmodelle (LLMs) und multimodalen Modelle bildet. LLMs sind eigentlich Textvervollständiger, die während des Trainings anhand von Lückentexten lernen, die jeweils fehlenden Wörter in einem Satz vorherzusagen. Der Aufmerksamkeitsmechanismus lenkt den Fokus auf Begriffe in dem oder den vorhergehenden Sätzen, die für den Kontext entscheidend sind, etwa: „Heute Nachmittag möchte ich mir ein Fahrrad kaufen. Ich gehe zur Bank und …". In diesem Fall wäre „hebe Geld ab" eine sinnvollere Fortsetzung als „setze mich hin" oder „hebe sie hoch". Anders als bei der Objekterkennung benötigten die Entwickler dafür keine von Clickworkern oder gar Experten angefertigten Trainingsbeispiele mehr, sondern konnten sie automatisiert aus sämtlichen im Internet veröffentlichten Texten generieren. Ein Skript zerlegt diese einfach in einzelne Sätze und entfernt je ein oder mehrere Wörter – fertig ist eine riesige Sammlung an Trainings-Samples inklusive Ground Truth. Die Ground Truth ist das jeweils entfernte Wort.

Diese Strategie brachte zusammen mit der neuen Transformer-Architektur den Durchbruch im Bereich der generativen KI. Weil die Trainingsdaten nicht mehr eigens von Menschen verschlagwortet werden mussten, mutierte das Internet plötzlich zur theoretisch unerschöpflichen Quelle, aus der sich der Wissensdurst der Sprachgeneratoren stillen ließ.

Backpropagation forever?

Aber eines hielt sich hartnäckig: der Backpropagation-Algorithmus. Anders als häufig behauptet ist das Training mit Lückentexten kein unüberwachtes Lernen, sondern nur eine autonom ablaufende Variante des überwachten Verfahrens. Ins LLM fließt ein unvollständiger Satz rein, hinten fällt eine Vorhersage des fehlenden Wortes raus. Eine Fehlerfunktion vergleicht die Vorhersage mit der Ground Truth, per Backpropagation werden die Parameterwerte nachjustiert. Trainiert wird so lange, bis der Fehler minimal ist. Eine korrekte Bezeichnung für diesen Prozess ist „selbstüberwachtes Lernen" und dieses unterscheidet sich nach wie vor fundamental von der vielfältigen Art und Weise, wie sich der Mensch Wissen aneignet.

Trotz dieser Einschränkung zeigten bereits die ersten Transformer mit Attention-Mechanismus wie das von Google entwickelte BERT ein beeindruckendes Gespür für Grammatik und Semantik, beginnend mit deutlich besseren Übersetzungen, die den Kontext berücksichtigten, bis hin zu Inhaltszusammenfassungen und thematisch konsistenten Texten. Damit schien erwiesen, was viele sich erhofft hatten: Dass sich zwischen den Zeilen all der Terabytes an geschriebenem Text ein kompaktes Modell des Weltwissens verbirgt – und die LLMs in der Lage seien, dieses zu extrahieren und zu interpretieren.

In gewisser Weise gelang ihnen das auch. Geleitet von der Aufgabe, Sätze zu vervollständigen, bildet sich in den verborgenen Schichten ein Algorithmus heraus, der Wörter nach semantischen Zusammenhängen sortiert. Ein LLM baut gewissermaßen einen gewaltigen, mehrsprachigen Index, der Begriffe theoretisch in jeden nur möglichen Bezug zueinander setzen kann – sei es bezüglich eines Themas oder einer bestimmten Art, sich auszudrücken: lyrisch, sachlich, in Code oder mathematisch.

Dieser Index, der sogenannte Latent Space, ist ein sehr hochdimensionaler Raum, in dem die Wörter und Wortfragmente (Token), die in einem bestimmten Kontext häufig zusammen vorkommen, nahe beieinander liegen. Die Token liegen dort allerdings nicht als Buchstabenfolgen, sondern sind als Vektoren kodiert, sodass sich Ähnlichkeiten leichter berechnen lassen. Eine Dimension kann konkrete Merkmale repräsentieren wie Farbe oder Oberfläche eines Objekts, aber auch abstraktere Konzepte wie Verwandtschaftsbeziehungen, gesellschaftliche Normen und Zahlensysteme.

Eine der aufregendsten Entdeckungen war, dass solche Netze mit wachsender Zahl an Parametern und Trainingsdaten anscheinend „Emergent Abilities" entwickelten: neue Fähigkeiten, die wie ein Sprung auf die nächste kognitive Ebene wirkten, etwa mehrstelliges Addieren, Subtrahieren und Multiplizieren oder die Fähigkeit, Fragen wahrheitsgemäß zu beantworten. Solche Aufgaben können kleinere LLMs nicht besser lösen als der Zufall, ab einer bestimmten Schwelle (Parameterzahl, Trainings-FLOPs) steigt

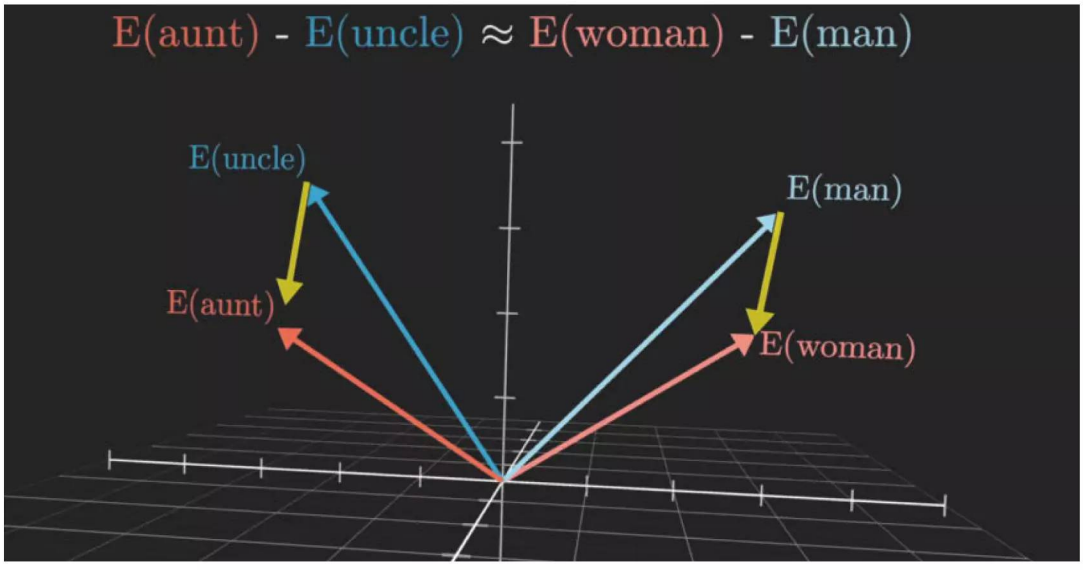


Bild: 3blue1brown

Im Latent Space liegen Token nach allen möglichen Merkmalen sortiert und als Vektoren codiert vor. Deshalb kann man mit den Eigenschaften sogar rechnen, etwa mit dem Unterschied zwischen männlich und weiblich (gelber Vektor).

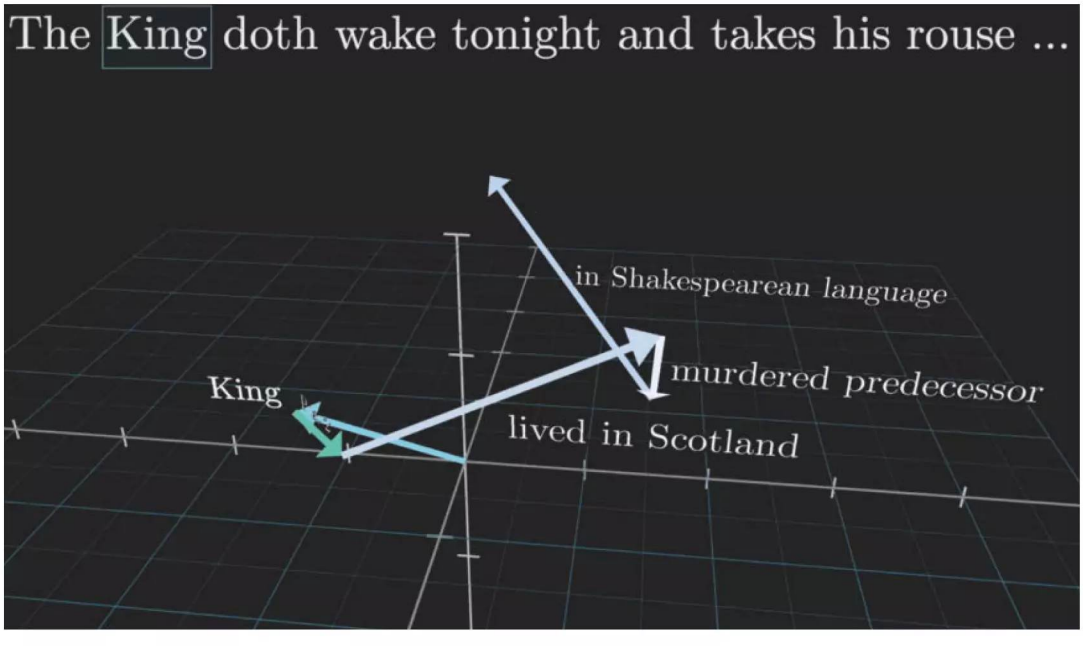


Bild: 3blue1brown

Schreiben bedeutet für ein Sprachmodell: eine Eingabe (Textprompt) in den Latent Space kodieren und zu diesen Codes passende Sätze bilden. Für die Vorhersage des nächsten Wortes nutzt es das Vokabular, das in der Nachbarschaft zu finden ist.

die Performance signifikant und relativ steil an. Eine gewisse Faktentreue stellte sich beispielsweise erst ab 280 Milliarden Parametern ein.

Evolutionssprünge

Anfangs funktionierte die Skalierungsstrategie wie am Schnürchen. Das 2018 veröffentlichte GPT-1 hatte 117 Millionen Parameter, verteilt auf 12 Layer und wandelte Token in 768-dimensionale Vektoren. GPT-2 kam mit 1,5 Milliarden Parametern, 48 Ebenen und produzierte 1600-dimensionale Vektoren. Mit dem 2020 veröffentlichten GPT-3 nochmal ein Vielfaches: 175 Milliarden Parameter, 96 Layer und ein Latent Space mit 12.288 Dimensionen. Seitdem gibt OpenAI keine konkreten Zahlen mehr über seine Modelle heraus und nur spärliche Informationen über die verwendete Architektur und Trainingsmethoden. Ziemlich sicher ist, dass GPT-4 und seine Nachfolger eine Mixture-of-Experts-Architektur (MoE) mit geschätzt einer Billion Parametern nutzen, Details siehe unten.

Forscher der Brown University fanden in Experimenten mit den frühen GPT-Versionen heraus, dass die Neuronen in den ersten Schichten eher naheliegende Muster kodieren, also etwa, dass auf „Donald“ häufig „Trump“ folgt oder auf „c’t Magazin für“ fast zwangsläufig „Computertechnik“. Ab der zwanzigsten Schicht kodierten sie auch abstraktere Zusammenhänge, sodass sie zum Beispiel den Ländern dieser Welt ihre Hauptstädte zuordnen konnten. Doch es ist das eine, aus Texten Faktenwissen zu ziehen und es nach den unterschiedlichsten, auch abstrakten Kriterien zu sortieren, und das andere, dieses zu verallgemeinern und auf neue Kontexte zu übertragen.

Deshalb hatten die lediglich aus Texten gespeisten Modelle auch enorme Probleme mit dem Rechnen und Zählen. Man kann davon ausgehen, dass von den einfachsten bis hin zu den komplexesten mathematischen Formeln und Rechnungen alles an Wissen in den Trainingsdaten vorhanden war. Aber offensichtlich genügte das nicht, um ein allgemeines Zahlenverständnis zu entwickeln oder gar die Regeln der höheren Mathematik daraus abzuleiten. Auch die Fähigkeit, Ursache und Wirkung zu verstehen oder Lösungen für komplexere Probleme zu entwickeln, ließ sich nicht einfach mit einer schieren Masse an Trainingsdaten oder noch größeren Netzen „erkaufen“. Zumal bereits so ziemlich das gesamte vom Menschen verfasste Textmaterial häppchenweise an LLMs verfüttert worden sein dürfte.

Dass es der Mensch mit so viel weniger Material- und Energieeinsatz schafft, liegt – wie Geoffrey Hinton bereits vor Jahren konstatierte – unter anderem daran, dass er offenbar effizientere Lernstrategien im Köcher hat als das langwierige Backpropagation-Verfahren. Diverse Studien aus der Hirnforschung halten die stereotypen Schleifen aus Feedback und Nachjustieren für biologisch nicht plausibel und auch der KI-Pionier Yann LeCun gibt sich überzeugt davon, dass es etwas Besseres und Zielführenderes geben muss.

Forscher der Universität Oxford haben immerhin ein Modell entwickelt, das versucht zu simulieren, wie das menschliche Gehirn zum ressourcensparenden Blitzmerker wird: Prospective Configuration. Demzufolge versetzt das Gehirn seine Neuronen situationsabhängig bereits vorab in einen Zustand, der eine korrekte Vorhersage wahrscheinlicher macht. Dazu erhöht es die Aktivität der vermutlich benötigten Neuronen entsprechend, trifft seine Vorhersage und verändert erst dann die Stärke der Synapsen. Das Lernen läuft also diametral zum Backpropagation-Verfahren, das mit inaktivem Neuronengeflecht beginnt.

Erste Experimente zeigten vielversprechende Ergebnisse, insbesondere beim Lernen aus wenigen Trainingsbeispielen, dem kontinuierlichen Lernen sowie der Anpassung an sich verändernde Bedingungen. Den Forschern zufolge ist es auch biologisch plausibler: So können natürliche neurale Aktivitätsmuster in verschiedenen Lernexperimenten durch Prospective Configuration erklärt werden, nicht aber durch Backpropagation – sowohl beim Menschen als auch beim Tier.

Episoden statt Wörter

Gerade erinnere ich mich, wie überrascht ich war, als ich nach langen Jahren mal wieder ein Schwimmbad betrat. Ich mag Schwimmbäder nicht sonderlich, aber mit Kindern geht man ja gerne hin. Wir holten die unvermeidlichen Pommes, und da war er wieder: dieser Schwimmbadgeruch aus meiner Kindheit, diese Mischung aus Chlorwasser auf nassem Teer in der Sommerhitze, gemischt mit Frittierfett. Sofort sah ich die Imbissbude des Augsburger Bärenkellerbades vor mir, die ganze Wiese, das faszinierende Tragluft-Hallenbad. Erstaunlich, was sich das Gehirn so merkt.

In diesem sogenannten episodischen Gedächtnis liegt ein fundamentaler Unterschied zwischen Mensch und Maschine, also wie der Mensch Infor-

mationen kontinuierlich, mit allen Sinnen, emotional oder rational gesteuert aus seiner Umwelt aufnimmt. Die Details zu seinen Erlebnissen speichert er meist nicht isoliert, sondern in einem größeren Zusammenhang: zum Beispiel auch die Erinnerung an ein schönes Abendessen mit Freunden. Es genügt ein kleiner Hinweis wie etwa die Frage, wer alles dabei war, und schon erscheint die gesamte Szene vor dem inneren Auge: der Ort, die angeregten Gespräche, wie gut das Essen geschmeckt hat, der Anlass des Treffens et cetera. An diesem Prozess sind verschiedene Gehirnregionen beteiligt: der Hippocampus, der die Episoden aufnimmt und abruft, sowie der Neokortex, in dem sie dauerhaft gespeichert werden.

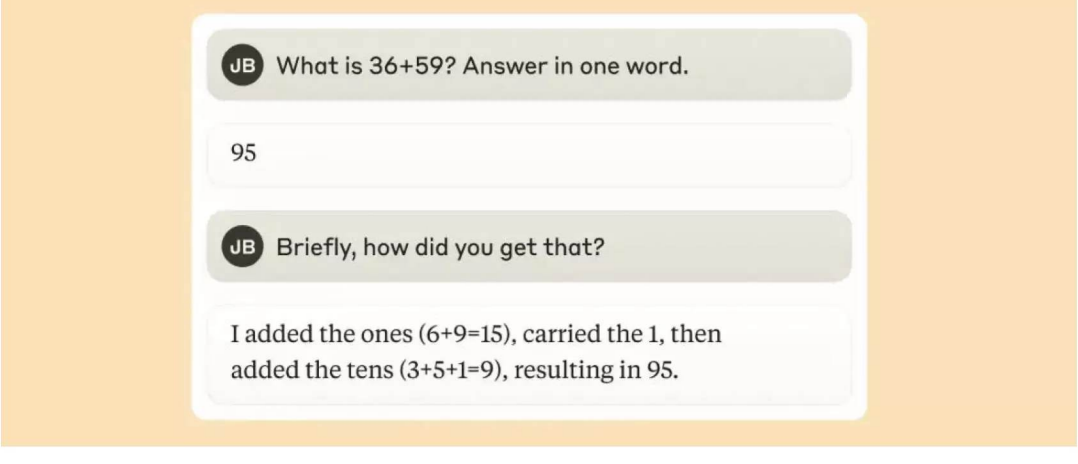
Neurowissenschaftler sind sich einig, dass dieses episodische Gedächtnis, das sowohl langsames als auch schnelles Lernen ermöglicht, die Grundvoraussetzung dafür ist, dass der Mensch sich rasch auf veränderte Situationen oder gar gänzliches „Neuland“ einstellen kann. Dass er kontinuierlich lernen, seine Wissensbasis erweitern sowie frühere Überzeugungen hinterfragen und gegebenenfalls verwerfen kann. Das episodische Gedächtnis hat bisher noch keine Entsprechung in einem der großen Sprachmodelle.

Experten halten eine solche Funktion aber für essenziell, um LLMs zu alltagstauglichen Agenten weiterzuentwickeln, sprich: zu Systemen, die in komplexen Umgebungen überlegt und sinnvoll handeln, auch über einen längeren Zeitraum hinweg. Man bezeichnet diese Fähigkeit, ein unbekanntes Problem auf Anhieb zu erfassen und zu lösen, auch als Single-Shot-Learning. Im menschlichen Alltag können das ganz banale Dinge sein, etwa eine Tür oder Schublade mit ungewöhnlichem Griff oder Schließmechanismus zu öffnen – für Roboter eine echte Herausforderung.

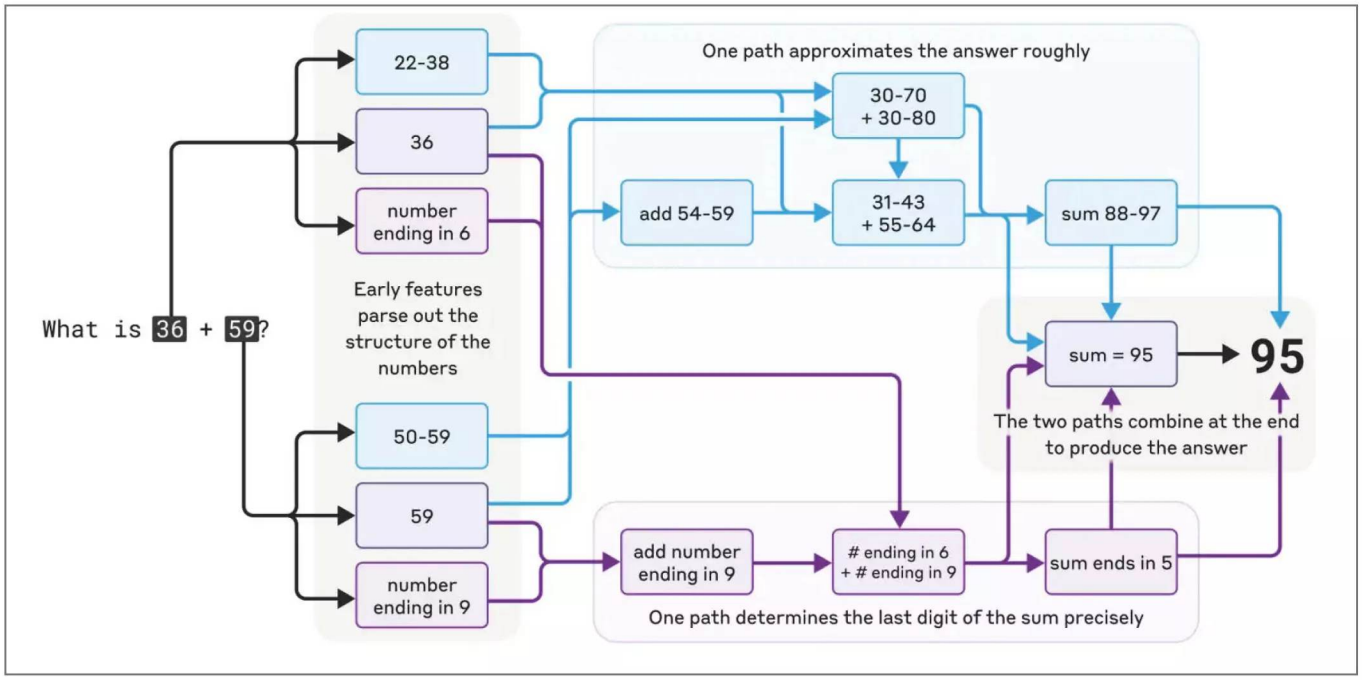
1, 2, 3, viele

Zentral ist das kontinuierliche Lernen auch beim Zahlenverständnis. Menschen und viele Tiere wie Affen, Krähen und Bienen haben einen angeborenen Sinn für kleine Zahlen bis vier, die jeweils fest mit einzelnen Neuronen verknüpft sind. Den Umgang mit größeren Zahlen verinnerlicht der Homo sapiens aber erst im Laufe seines Lebens, in einer Mischung aus schulischem und beiläufigem Lernen. Sobald sich Kulturen oder einzelne Individuen mathematische Fähigkeiten aneignen, nehmen sie ihre Umwelt stärker unter solchen formalen oder numerischen Aspekten wahr. So lernen sie beiläufig, die abstrakten Regeln auf konkrete Objekte und ihren Alltag anzuwenden: zum Beispiel, dass man außer Äpfeln und Birnen auch Buchstaben, Grashalme, Noten, Töne oder was auch immer zählen kann. Das Gesamte zerlegt der Mensch dafür einfach, ohne darüber nachdenken zu müssen, in seine Einzelteile.

Große Sprachmodelle haben oder hatten ihre Mühe mit dem Zählen und beim Rechnen mit sehr



Als Erklärung lieferte das Reasoning-Modell indes den in der Schule vermittelten Rechenweg.



Wenn ein Mensch seine Entscheidung begründet, kann er lügen oder die Wahrheit sagen. Sprachmodelle sind sich ihrer internen Prozesse schlicht nicht bewusst, wie Anthropic am Beispiel einer einfachen Rechnung zeigt. Intern löst Claude Sonnet die Aufgabe mit einer Kombination aus Rechnen und Schätzen.

großen oder untypischen Zahlen, die in den Trainingsbeispielen eher selten vorkommen. Ein Grund dafür ist, dass die ersten LLMs schlichtweg weder ein Zahlenverständnis entwickelt noch das Rechnen gelernt hatten, sondern ein Schema, für bestimmte, nur in Details variierende Aufgabentypen die plausibelste Lösung vorherzusagen. Ein weiteres Problem liegt in der Tokenisierung der Eingabedaten, bei der Sätze, Wörter und auch Zahlen in Fragmente aufgeteilt werden. Die Standardverfahren sind darauf optimiert, korrekte Sätze zu bilden – und nicht aufs Rechnen oder Zählen.

Menschlicher Feinschliff

Die Hoffnung oder vielmehr das Versprechen, dass ein Sprachmodell mit schierem Einsatz von Ressourcen automatisch zum Weltmodell wird, hat sich bis dato also nicht erfüllt. Um grundlegende Schwächen auszubügeln, Wissenslücken zu schließen und die Chatbots auf Faktentreue und eine angenehme Gesprächskultur einzuschwören, brauchte man doch wieder den Menschen als Experten und Lehrer. Deshalb schließt sich an das selbstüberwachte Grundlagentraining ein sehr aufwendiges, überwachtes Finetuning mit händisch gelabelten Trainingsdaten an: das Reinforcement Learning from Human Feedback (RLHF).

Zur Gewinnung von Trainingsdaten vergleichen menschliche Prüfer Sprachmodell-Ausgaben zu diversen Prompts mit jeweils von Experten verfassten Antworten und bewerten, wie nahe die LLM-Ergebnisse dem Ideal kommen. Mithilfe solcher Paarungen aus generierten Texten und menschlichem Ranking lernt ein weiteres großes Sprachmodell, Texte nach ähnlichen Kriterien wie die menschlichen Gutachter einzuschätzen. Fortan kann es als maschinelle Feedback-Schleife dienen, die die Antworten des Hauptmodells prüft, verwirft oder verbessert.

Ganz neu sind Reinforcement-Learning-Strategien, die LLMs dazu zwingen, komplexe Aufgaben sinnvoll in kleinere Schritte aufzuteilen und formal strukturierte Antworten auszugeben, sodass der Mensch nachvollziehen kann, wie das Modell zu seiner Lösung gekommen ist und an welcher Stelle ihm möglicherweise ein Fehler unterlief. Bei dieser sogenannten Test-time-compute-Methode handelt es sich nicht um ein Finetuning, sondern um ein Machine-Learning-Verfahren, das zur Laufzeit angewendet wird, um die schlechte Bilanz der LLMs in den einschlägigen Mathe- und Logik-Benchmarks zu polieren.

Alle aktuellen Reasoning-Modelle nutzen Test-time compute. Die DeepSeek-Entwickler erklären in einem Aufsatz sogar, wie sie es bei ihrem R1-Modell umsetzen: mit selbstüberwachtem Reinforcement Learning, bei dem iterativ eine Menge möglicher Antwortvarianten generiert und die zielführendsten Ideen belohnt werden. Mit einem ähnlichen Ansatz schaffte es einst AlphaGo, sich beim Brettspiel Go vom Vorbild menschlicher Strategien zu lösen und eigene, überraschende Kombinationen zu finden.

Das R1-Belohnungssystem ist ebenfalls ein trainiertes Modell: Anhand von Musterlösungen für Programmierprobleme, anspruchsvolle Mathematikaufgaben oder Logikrätsel hat es gelernt, verständliche und gut strukturierte Schritt-für-Schritt-Antworten sowie korrekte Lösungen mit höheren Scores zu bewerten. Auf dieser Grundlage wählt es aus mehreren gesampelten Varianten die vielversprechendsten aus und animiert das Modell zudem, in ähnlicher Qualität fortzufahren. Es ist also eine Optimierungsmethode, die für mathematisch-naturwissenschaftliche Aufgaben mit nur einem korrekten Ergebnis sehr gut funktioniert, bei offeneren Fragestellungen allerdings weniger.

Und der Rechenaufwand ist enorm. Zwar orientieren sich die Entwickler an der dynamischen Arbeitsweise des menschlichen Gehirns, das einfache Fragen rasch beantwortet und erst für komplexe Probleme mehr Ressourcen aktiviert, sprich: nachdenkt. Allerdings muss der Mensch nicht in Brute-Force-Manier wieder und wieder enorme Mengen an

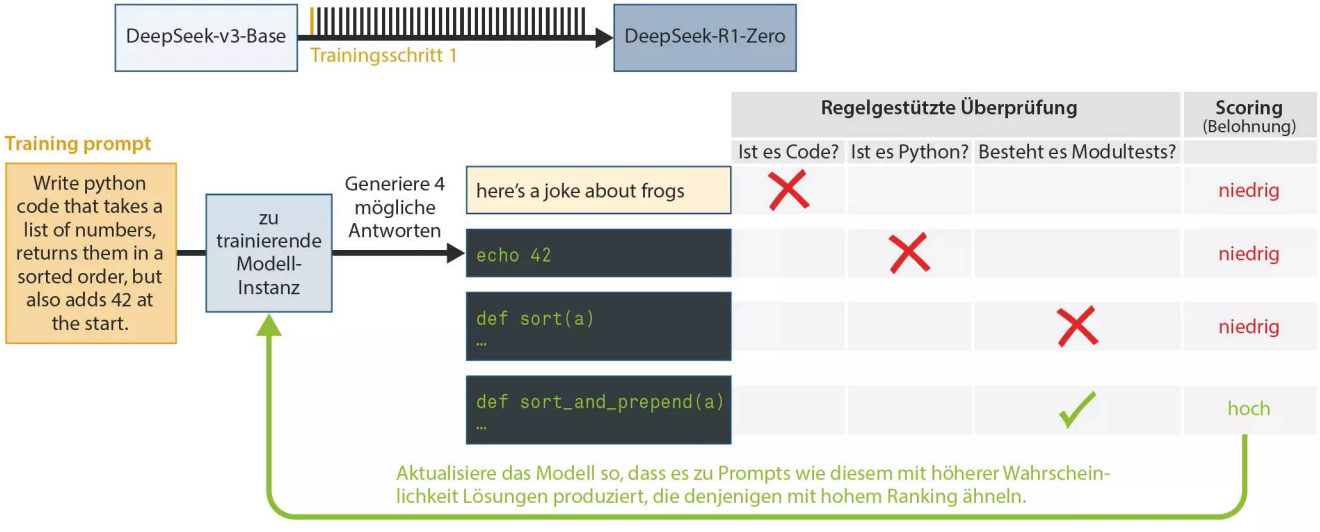
Was menschliches von maschinellem Lernen unterscheidet

	Mensch	künstliche Intelligenz
Art des Lernens	erfahrungsbasiert, adaptiv, kontinuierlich	datenbasiert, oft offline in Trainingsphasen
Lernmechanismus	synaptische Plastizität	Backpropagation und Gradientenabstieg
Flexibilität	äußerst flexibel (Transferlernen aus wenigen Beispielen)	aufgabenorientiert; oft beschränkt auf Trainingsverteilung
Datenbedarf	wenige Beispiele nötig (Few-Shot-/One-Shot-Learning)	mehrere hundert bis zigtausende Beispiele notwendig (je nach Aufgabe)
Anpassungsfähigkeit	laufende Anpassung in Echtzeit möglich (kontinuierliches Lernen)	Anpassung nur durch Nachtraining oder Finetuning
Verallgemeinerung	sehr stark (auch auf unbekannte Situationen)	abhängig von Trainingsdaten und Prompt-Engineering
Fehlerverarbeitung	intuitive Korrektur, Lernen aus Fehlern und Exploration	Fehler werden durch Updates in Trainingsphasen behoben
Gedächtnisstruktur	semantisches, episodisches und prozedurales Gedächtnis	parametrisches Wissen, gegebenenfalls ergänzt um Retrieval-Systeme (RAG)
Energiebedarf	sehr energieeffizient (~20 W Verbrauch im Ruhezustand)	sehr energieintensiv (GPU: bis 700 W; KI-Rechenzentrum: 100 bis 1000 MW)

Quelle: Jay Allamar

Wie DeepSeek das Reasoning lernt

DeepSeek R1 verfasst Schritt-für-Schritt-Lösungen oder auch validen Code, weil ein Reinforcement-Learning-System es dafür belohnt. Zuerst generiert es sehr viele mögliche Antworten zu einem Prompt, die auf Konformität zu einigen wenigen Regeln geprüft werden. Dann werden die Modellparameter so aktualisiert, dass das Sprachmodell hoch bewertete Antworten mit höherer Wahrscheinlichkeit ausgibt als solche mit niedrigem Score.



möglichen Antworten generieren; sein Gehirn hinterfragt die eigenen Gedankengänge zielgerichteter.

Messen impossible

Wie genau sich Menschen Wissen und Fähigkeiten aneignen, warum sie sich manches merken und anderes nicht und was genau jedes Individuum dazu motiviert oder davon abhält, einer Sache auf den Grund zu gehen: Die Mechanismen dahinter sind bisher nur ansatzweise erforscht, unter anderem weil man die Gehirnaktivität lebender, gesunder Menschen nur mit nicht-invasiven Methoden messen kann. Eine weit verbreitete Methode ist die funktionelle Magnetresonanztomographie (fMRT); in der Literatur gebräuchlicher ist die englische Abkürzung fMRI. Genauere Messwerte auf Basis einzelner Neuronen liefern invasive Verfahren mit im Cortex platzierten Elektroden oder Nadeln; entsprechende Studien werden in der Regel aber nur mit Epilepsie-Patienten durchgeführt, die bereits ein Implantat haben.

Auch Large Language Models lassen sich nicht so einfach in die Synapsen schauen, unter anderem weil das Gros der High-End-Modelle proprietär ist und daher der unabhängigen Forschung nicht zur Verfügung steht. Viele Studien über die im LLM simulierten Denkprozesse stützen sich daher auf die frühen Versionen der OpenAI-Modelle GPT-1 bis GPT-3 oder kommen direkt aus den Research-Abteilungen der Hersteller. Einige interessante Erkenntnisse veröffentlicht immerhin Anthropic über die internen Prozesse seines Sprachmodells Claude Sonnet.

Quintessenz: Es ist vertrackt, denn anscheinend kann man sich selbst bei einem Reasoning-Modell nicht darauf verlassen, dass es erklärt, wie es tatsächlich zu seinem Ergebnis gekommen ist. Es hat ja lediglich gelernt, dem Menschen einen logisch nachvollziehbaren Lösungsweg zu präsentieren. (atr) c't

Literatur

[1] Andrea Trinkwalder, Netzgespinste, Die Mathematik neuronaler Netze: einfache Mechanismen, komplexe Konstruktion, c't 6/2016, S. 130

Wissenschaftliche Aufsätze, Simulationen, Videos:
ct.de/w981

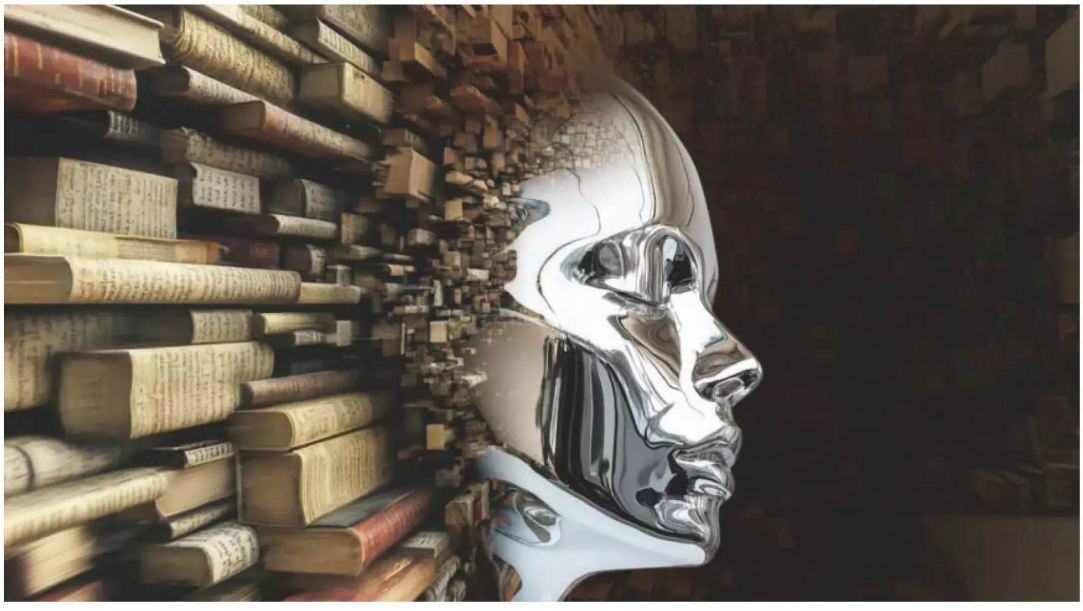


Bild: KI, Collage c't

KI mit Fachwissen ausstatten, Teil 1

Das in Sprachmodellen wie ChatGPT gespeicherte Wissen ist nach der Trainingsphase fixiert. Aktuelle Informationen oder Know-how für spezielle Anwendungsfälle muss man ihnen per Retrieval Augmented Generation erschließen.

Von **Sebastian Springer**

Sprachmodelle wie ChatGPT, Gemini und Llama werden mit großen Datenmengen trainiert. Ihr Wissen bleibt allerdings auf den Stand des Trainingskorpus beschränkt. Schon bei einfachen Fragen wie „Was gibt es morgen Mittag in der Kantine zu essen?“ passen die Sprachmodelle oder – schlimmer – sie halluzinieren, weil ihnen das notwendige Wissen fehlt.

Zwar können die großen kommerziellen Large Language Models (LLMs) mittlerweile das Internet als Quelle in ihre Antworten einbeziehen. Doch auch diese Fähigkeit hat ihre Grenzen – insbesondere, wenn es um persönliche oder firmeninterne Informationen geht, denn diese stehen oft nicht öffentlich zur Verfügung.

Die Modelle sind also ohne weiteres Zutun nicht für spezialisierte oder vertrauliche Kontexte geeignet. Doch es gibt verschiedene Ansätze, um LLMs mit Fachwissen auszustatten. Dieser Artikel skizziert diese Methoden und zeigt an einem einfachen Bei-

spiel, wie man die sogenannte Retrieval Augmented Generation (RAG) mit dem Framework LangChain umsetzt.

Das Sprachmodell neu erziehen

Man kann Sprachmodellen Fachwissen beibringen, indem man sie mit dem spezifischen Wissen neu trainiert, sie nachoptimiert oder sie per RAG mit externen Daten versorgt. Beim Training eines LLM fängt das Modell bei null an, ähnlich wie ein Kind, das Sprache und Wissen von Grund auf lernt.

Das Sprachmodell lernt also den großen allgemeinen Trainingsdatensatz, den es benötigt, um die Sprache überhaupt zu lernen. Zusätzlich fließen die domänenspezifischen Inhalte in sein Training mit ein. Es wird also auf einer enorm großen Datenmenge trainiert.

Der Trainingsprozess eines LLM ist rechenintensiv und entsprechend teuer. In der Regel ist aber die Menge der fachspezifischen Daten im Vergleich zum allgemeinen Trainingskorpus sehr klein. Das führt dazu, dass sie den Output kaum verändern. Man schießt mit einem Neutraining also mit Kanonen auf Spatzen.

Finetuning baut auf einem bereits vortrainierten Modell auf. Dabei trainiert man das Modell mit einem kleineren, spezielleren Datensatz weiter. Dieser Ansatz erinnert an das Anlernen eines Facharbeiters: Jemand, der bereits über Grundkenntnisse in einem Berufsfeld verfügt, lernt gezielt die Anforderungen für eine bestimmte Tätigkeit nach.

Das Finetuning ist allerdings ebenfalls aufwendig. Es benötigt speziell aufbereitete Daten und Fachwissen für die Implementierung. Und das Ergebnis ist selbst wieder ein starres Sprachmodell, das nicht mal schnell an eine geänderte Informationslage angepasst werden kann.

Bei der Retrieval Augmented Generation bleibt das LLM in seinem ursprünglichen Zustand. Es greift auf eine Wissensdatenbank oder andere spezialisierte Informationsquellen zu, während es seine Antworten generiert. Man kann sich das wie einen Experten vorstellen, der sich bei schwierigen Fragen auf ein Handbuch oder eine gut organisierte Sammlung von Notizen stützt. Der Experte muss nicht alles auswendig wissen, er kann gezielt nachschlagen und die Informationen dann auf sinnvolle Weise kombinieren.

Damit ist RAG vergleichsweise leichtgewichtig und flexibel, weil es keine Änderungen am Modell erfordert. Gleichzeitig lassen sich so aktuelle und sehr spezielle Informationen in Echtzeit nutzen, ohne das Modell aufwendig neu zu trainieren oder anzupassen: Der Experte bleibt gleich; man tauscht einfach das Handbuch aus, um im Bild zu bleiben.

Sprach-KI-Baukasten

Dieser Artikel beschreibt, wie Sie eine simple Version des Onlinedienstes ChatPDF bauen: eine Anwendung, die ein PDF-Dokument lädt und analysiert und der Sie anschließend Fragen zu dem PDF stellen können. Das Beispiel nutzt das Open-Source-Framework LangChain. Es wurde entwickelt, um Sprachmodelle in Anwendungen einzubinden. LangChain ist eine Art Werkzeugkasten. Er enthält eine Reihe von Funktionsblöcken, die man für die Verarbeitung von Texten und die Zusammenarbeit mit Sprachmodellen benötigt, zum Beispiel für Retrieval Augmented Generation. Diese Blöcke lassen sich zu sogenannten Chains verketten. Entwickler können damit komplexe Workflows aufsetzen.

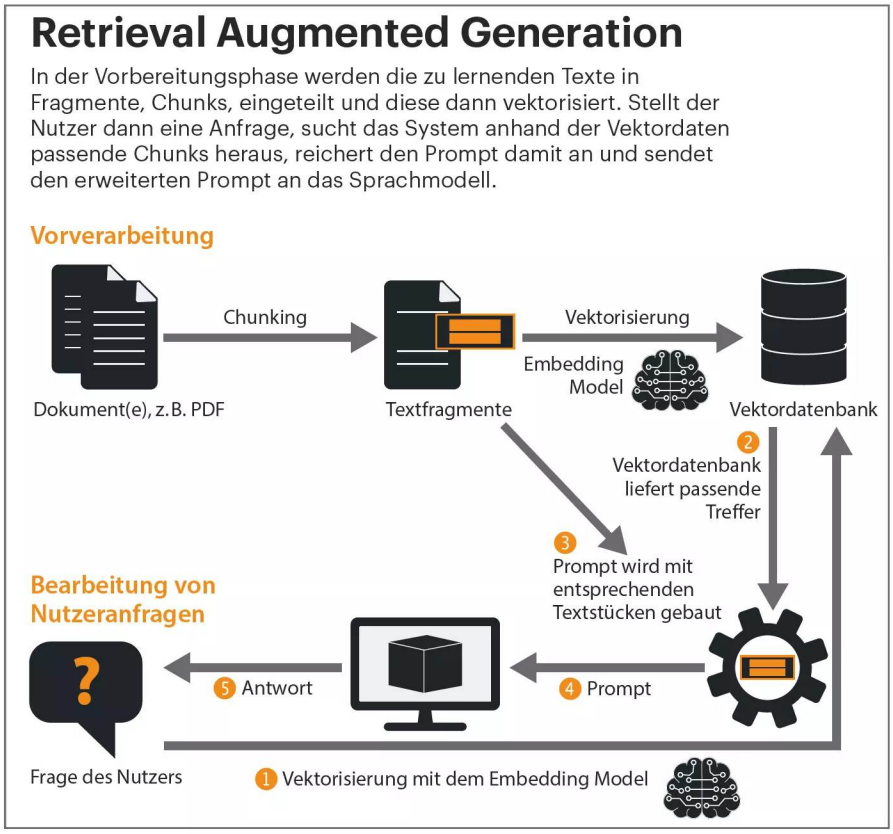
LangChain existiert in einer JavaScript- und einer Python-Variante und deckt damit die zwei aktuell am weitesten verbreiteten Programmiersprachen für KI-Applikationen ab. Die diesem Artikel zugrundeliegende JavaScript-Variante lässt sich in den verschiedensten Laufzeitumgebungen einsetzen, darunter Node.js, Cloudflare Workers, Next.js und sogar im Browser.

Das Beispiel setzt eine Node.js-Installation voraus. LangChain selbst lässt sich mit der Anweisung `npm add @langchain/core` installieren. Neben LangChain nutzt das Beispiel Ollama als Schnittstelle zu beziehungsweise Laufzeitumgebung für Sprachmodelle sowie die Vektordatenbank Milvus.

Texte mundfertig servieren

Die Beispielanwendung besteht aus kleinen, für RAG typischen Blöcken. Ein RAG-System durchläuft nicht alle Schritte in Echtzeit, während es seine Antworten generiert. Vielmehr muss der Entwickler einige Vorarbeiten erledigen. So muss das RAG-System die Daten zunächst aus externen Datenquellen einlesen.

Der nächste Arbeitsschritt nennt sich Chunking. Dabei teilt man die Daten in Blöcke auf, die das RAG-System gut weiterverarbeiten kann. Der dritte Schritt wandelt diese Chunks in numerische Vektoren um. Die Vektorisierung bildet Bedeutungen, Beziehungen und Ähnlichkeiten zwischen Wörtern oder Sätzen mathematisch ab. Das System speichert diese Vektoren schließlich in einer Datenbank.



Ablauf des RAG: Zunächst müssen die Inhalte aufbereitet werden, um für das Sprachmodell auf Abruf bereitzustehen.

Mithilfe der vorbereiteten Daten kann das RAG-System dann eine Nutzeranfrage mit dem allgemeinen Wissen des Sprachmodells und dem hinzugefügten Fachwissen beantworten. Zu einer Anfrage des Anwenders nutzt es die Ähnlichkeitssuche der Vektordatenbank, um passende Textstellen zu finden. Die RAG-Anwendung ergänzt den Prompt des Nutzers um die in der Datenbank gefundenen Textabschnitte, überträgt sie an das Sprachmodell und erhält daraus eine Antwort.

Die Daten können aus jeder beliebigen Quelle stammen. Üblich sind textbasierte Formate, zum Beispiel PDF- oder Markdown-Dateien, Webseiten oder Datenbanken. Der erste Codeausschnitt zeigt, wie eine Funktion in Node.js mit dem `PDFLoader` der LangChain-Bibliothek den Inhalt einer PDF-Datei in Text umwandelt . Alle Codeausschnitte und das Gesamtprojekt finden Sie unter ct.de/w6ff.

Das Ergebnis dieses Funktionsblocks ist der Inhalt einer PDF-Datei, deren Namen er als Argument beim Aufruf als JavaScript-Zeichenkette übergeben bekommt. Der Umfang der Datei ist nur durch die Ressourcen des Systems begrenzt. So sind auch Bücher mit über 1000 Seiten kein Problem für diese Umwandlung. Die `RunnableLambda`-Klasse sorgt dafür, dass LangChain die einzelnen Codeblöcke zu einer Sequenz oder Kette zusammenfügt und sie im Kontext der Gesamtapplikation ausführen kann. Die Funktion ist rasend schnell und benötigt bei einem Test auf einem MacBook Pro M1 für die Umwandlung eines Buches mit etwa 550 Seiten nur 1,1 Sekunden.

Klein schneiden

Das Chunking teilt die Textdaten der Quelle(n) in kleinere Segmente auf (Chunks, Stückchen). Der Grund hierfür ist, dass ein Sprachmodell nur über eine bestimmte Kapazität für Zeichen verfügt, die es pro Anfrage verarbeiten kann, den sogenannten Kontext.

Will man beispielsweise einen Chatbot bauen, der bei der Entwicklung moderner Web-Applikationen mit dem Framework React hilft, kommen schnell

```
import { RunnableLambda } from '@langchain/core/runnables';
import { PDFLoader } from '@langchain/community/document_loaders/fs/pdf';

const loadPDF = new RunnableLambda({
  func: async (file: string): Promise<Document> => {
    const loader = new PDFLoader(file, {
      splitPages: false,
    });
    const docs = await loader.load();
    return docs[0];
  },

});
```

LangChain stellt eine Reihe von Loader-Klassen für den Import verschiedener Dokumenttypen bereit, zum Beispiel `PDFLoader` für PDF-Dokumente.

einige Hundert Kilobyte an Material zusammen: die JavaScript- und TypeScript-Spezifikationen, die Offline-Variante des Mozilla Developer Networks und die gesamte React-Dokumentation. Zusätzlich dazu könnte die Applikation noch auf verschiedene Fachbücher zu diesem Themengebiet zugreifen.

Die JavaScript-Spezifikation allein mit ihren rund 267.000 Wörtern und knapp 2 Millionen Zeichen ist schon deutlich größer als der Kontext vieler Sprachmodelle. Das Kontextfenster bei LLama 3 oder GPT-4o etwa fasst 128.000 sogenannter Token, Gemini 2.0 Pro kann bis zu 2 Millionen Token verwenden. Ein Token kann ein Wort, ein Teil eines Wortes, ein Satzzeichen oder ein Leerzeichen repräsentieren.

Token lassen sich also nicht 1:1 in eine Anzahl von Wörtern umrechnen. Eine Daumenregel sagt aus, dass ein Token etwa drei Viertel eines Wortes entspricht. Ein Kontext mit einer Größe von 128k Token entspricht also rund 96.240 Wörtern. Geht man davon aus, dass eine DIN A4-Seite Text etwa 500 Wörter umfasst, sind das 192 Seiten.

Verarbeitet ein Sprachmodell nur die Chunks, die für den aktuellen Prompt relevant sind, muss es nicht mehr die gesamten verfügbaren Dokumente einlesen. Das versetzt das System in die Lage, eine zielgerichtete Antwort zu erzeugen. Außerdem arbeitet das Gesamtsystem mit kleineren Textabschnitten deutlich schneller und damit ressourcenschonender.

Nutzt man eines der großen Sprachmodelle via API, das die Betreiber nach der übertragenen Datenmenge abrechnen, spart man so auch Geld.

Für das Chunking gib es verschiedene Strategien. Die einfachste Variante ist das zeichenbasierte Chunking. Dabei unterteilt das System den Text in gleich große Abschnitte mit einer festen Zeichenlänge. Diese Strategie ignoriert allerdings die Textinhalte. Der Text wird einfach abgeschnitten, egal ob mitten in einem Wort, Satz oder Absatz.

Besser sind Strategien, die sich an den Textinhalten orientieren. Nutzt das System beispielsweise Absätze als Begrenzung, ist die Wahrscheinlichkeit hoch, dass es zusammengehörende Informationen in einem Chunk unterbringt. Neben diesen Strategien existieren weitere, noch deutlich komplexere, die beispielsweise Chunks anhand der Semantik eines Textes bilden. Hier kommen Werkzeuge zur Sprachverarbeitung zum Einsatz, die die Bedeutung der Input-Texte analysieren und mit diesen Informationen die Chunks bilden.

Eine weitere Maßnahme, die Qualität der Chunks zu verbessern, funktioniert komplementär zu den grundlegenden Chunking-Strategien: das Überlappen der Chunks. Man zertrennt die Texte also nicht scharf, sondern lässt die Chunks 20 bis 30 Prozent überlappen. Das kann helfen zu vermeiden, dass das Chunking Zusammenhänge zerteilt.

Wohldimensionierte Häppchen

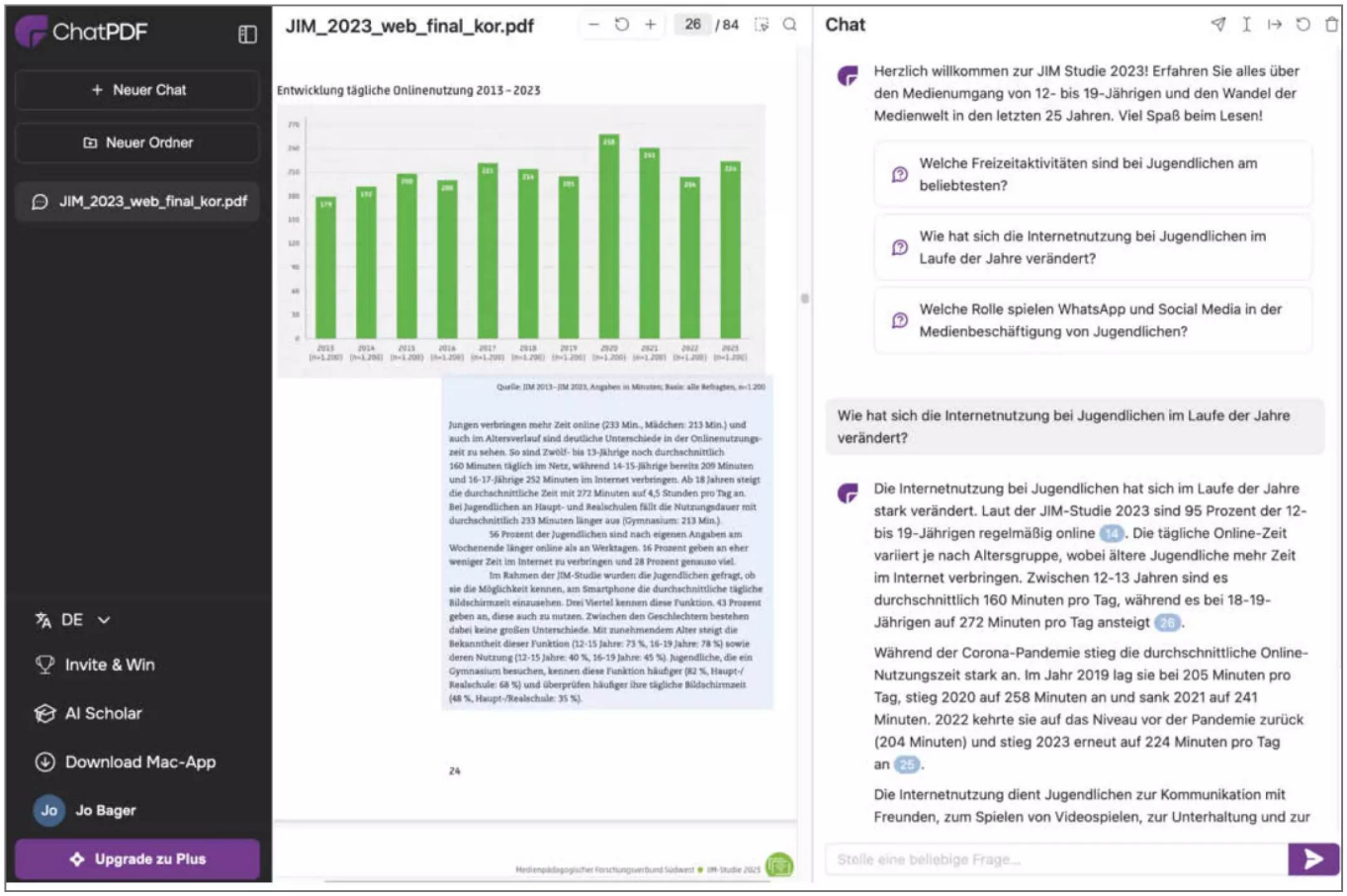
Der zweite Codeausschnitt zeigt ein Beispiel für ein einfaches zeichenbasiertes Chunking mit der `RecursiveCharacterTextSplitter`-Klasse aus der LangChain-Bibliothek. Der Konstruktor der Klasse akzeptiert unter anderem die `chunkSize`, also die gewünschte Größe der Textblöcke, sowie einen `chunkOverlap`, also die Überlappung der Chunks. Beide Werte geben die jeweilige Anzahl von Zeichen an.

Die Größe der Chunks hängt etwas vom Anwendungsfall ab. Größere Chunks stellen sicher, dass der Kontext im Text erhalten bleibt, sie sind jedoch auch langsamer und kosten mehr Ressourcen wie Rechenzeit und Speicher bei der Auswertung. Kleinere Chunks sind schneller und beim Ressourcenverbrauch günstiger, bergen jedoch die Gefahr, dass Informationen durch die vielen Textschnitte verloren gehen. Im Beispiel liegt die Chunkgröße bei 1000 Zeichen mit einer Überschneidung von 200 Zeichen.

Der `RecursiveCharacterTextSplitter` von LangChain versucht, größere Textblöcke beim Aufteilen intakt zu halten. Überschreitet ein Satz das Zeichenlimit, ver-

```
import { RecursiveCharacterTextSplitter } from '@langchain/textsplitters';
import { RunnableLambda } from '@langchain/core/runnables';

const chunkSize = 1000;
const chunkOverlap = 200;

const splitText = new RunnableLambda({
  func: async (document: Document) => {
    const splitter = new RecursiveCharacterTextSplitter({
      chunkSize,
      chunkOverlap,
    });
    const texts = await splitter.splitText(document.pageContent);
    return { texts, metadata: flattenObject(document.metadata) };
  },

});
```

Mit der `chunkSize` und dem `chunkOverlap` der `RecursiveCharacterTextSplitter`-Klasse gibt man die Größe sowie die Überlappung der Chunks in Zeichen an.



PDF-Datei hochladen und anschließend einem Chatbot Fragen dazu stellen: ChatPDF ist eine populäre, jedermann zugängliche RAG-Anwendung.

schiebt ihn die Bibliothek in den nächsten Chunk. Das führt dazu, dass die Chunks gegebenenfalls etwas kleiner als die gewünschte Zeichenanzahl sind. Außer den Text-Chunks extrahiert diese Funktion die Metadaten aus dem Dokument und reicht sie an das folgende Element der Kette weiter, sodass nicht nur der Text, sondern auch weitere Informationen wie der ursprüngliche Dateiname verfügbar sind.

LangChain enthält noch weitere TextSplitter-Implementierungen wie beispielsweise den `CharacterTextSplitter`, der keine Rücksicht auf die Textstruktur nimmt und gleichmäßig große Chunks produziert, oder den `TokenTextSplitter`, der nicht mit regulären Zeichen und Wörtern, sondern auf Basis von Token arbeitet.

Eine Spezialität der `RecursiveCharacterTextSplitter`-Klasse ist, dass sie verschiedene Sprachen beziehungsweise Textstrukturen wie Markdown, Latex, HTML, Python oder JavaScript unterstützt. Dabei berücksichtigt der TextSplitter die speziellen Trennzeichen der jeweiligen Sprache zur Aufteilung des Texts.

Die Text-Chunks sind die Grundlage für die nächste Phase im Vorbereitungsprozess des RAG-Systems: Sogenannte Embeddings wandeln die Chunks in Vektoren und speichern sie anschließend in einer speziellen Datenbank. Wie das vonstattengeht und wie man die Vektoren und die Chunks schließlich nutzt, um ein Sprachmodell mit Spezialwissen auszustatten, erklärt der zweite Teil dieses Artikels, den Sie auf Seite 32 finden. (jo) c't

Weitere Informationen zu LangChain:
ct.de/w6ff

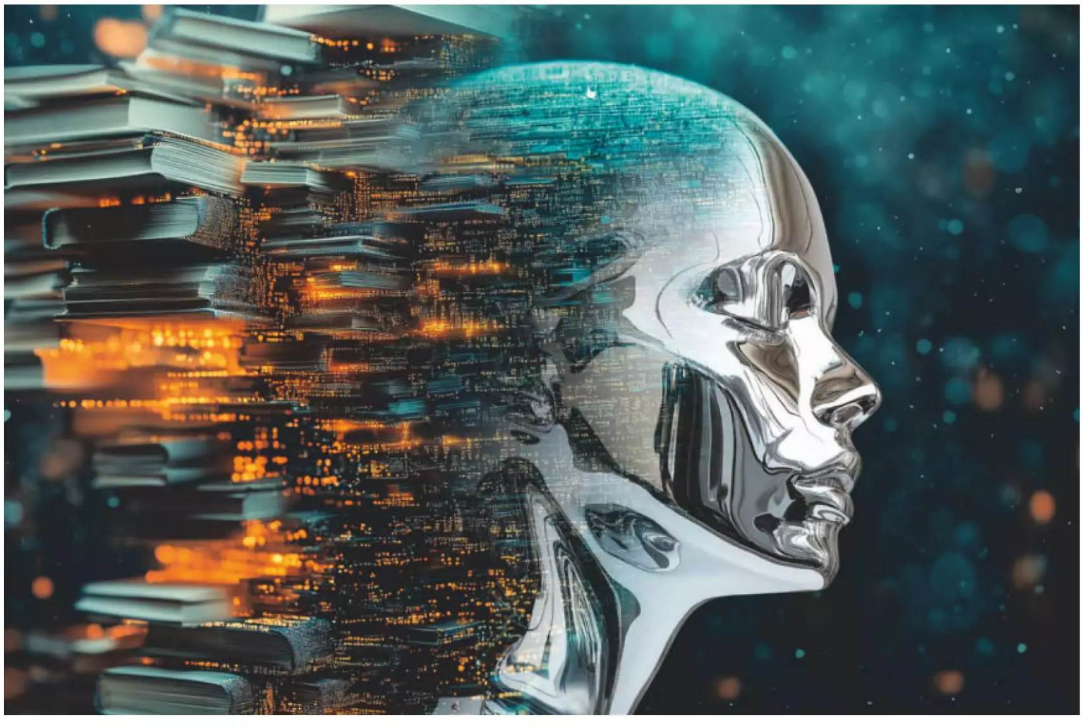


Bild: KI, Collage c't

Sprachmodelle mit RAG feintunen, Teil 2

Sprachmodelle können auf spezielles Wissen zugreifen, wenn man es ihnen häppchenweise im Prompt mitserviert. Die Kunst ist es dabei, aus einem großen Wissensfundus die richtigen Häppchen auszuwählen.

Von **Sebastian Springer**

Der erste Teil dieses Artikels hat gezeigt, wie Retrieval Augmented Generation (kurz RAG) grundsätzlich funktioniert und die ersten Schritte mit dem Framework LangChain erklärt. Zunächst liest ein RAG-System die benötigten Daten ein. Im zweiten Schritt teilt es sie in Häppchen auf, sogenannte Chunks. Stellt der Nutzer dem RAG-System eine Frage, wählt es einige zur Frage passende Chunks aus. Damit erweitert es die Anfrage, die es anschließend an das Sprachmodell stellt. Auf diese Weise erhält das Sprachmodell Kontext für seine Antwort.

Doch woher weiß das RAG-System, welche Chunks am besten zur Anfrage passen? Computer

können Texte ja nicht verstehen. Sie benötigen eine numerische Darstellung, um Muster oder Ähnlichkeiten zu erkennen. Vektorisierung erzeugt eine solche Darstellung, die maschinell besser verarbeitet werden kann. Für die Suche nach passenden Chunks nutzt das RAG-System dann eine vektorbasierte Ähnlichkeitssuche.

Vektorisierung wandelt Wörter oder Sätze in Vektoren um. Aus einem Satz wie: „Die Katze sitzt auf der Matte“ entsteht dabei ein Vektor wie dieser: [0.12, -0.34, 0.98, ...] Jeder Wert des Vektors steht für ein bestimmtes Merkmal des Satzes, etwa die Bedeutung einzelner Elemente oder den Kontext. Würde man den Satz: „Der Hund sitzt auf der Matte“ vektorisieren, würde sich sein Vektor wenig von dem des ersten Satzes unterscheiden, weil beide Sätze gleich aufgebaut sind und ähnliche Bedeutungen haben. Der Vektor von: „Martin fuhr mit seinem Maserati zu schnell in die Kurve hinein“ würde sich dagegen deutlich von beiden unterscheiden.

Modellkunde

Vektoren mit einer geringen Anzahl von Dimensionen (bis etwa 100) eignen sich lediglich für einfache Anwendungsfälle, da sie nicht in der Lage sind, komplexe Zusammenhänge abzubilden. Dafür sind sie sehr performant und liefern schnell Ergebnisse bei der Suche. Hochdimensionale Vektoren mit 1000 oder mehr Dimensionen eignen sich für komplexe Sachverhalte. Der Nachteil dieser Leistungsfähigkeit: hoher Speicherbedarf und hohe Rechenintensität.

Die mathematische Darstellung von Daten in Form von Vektoren nennt man auch Embeddings. Zum Umwandeln von Texten in Embeddings dienen speziell für diesen Zweck trainierte Sprachmodelle. Einfache Embedding-Modelle wie GloVe oder Word2Vec erzeugen Vektoren für einzelne Wörter, unabhängig von ihrem Kontext, und arbeiten mit 50 bis 300 Dimensionen.

Komplexere Modelle wie BERT, RoBERTa, der universal-sentence-encoder von Tensorflow und das kommerzielle text-embedding-3-large von OpenAI berücksichtigen auch den Kontext von Wörtern. Sie nutzen Hunderte Dimensionen, text-embedding-3-large sogar bis zu 3072 Dimensionen. Die Beispiel-Applikation nutzt das „nomic-embed-text“-Modell, um aus den Text-Chunks Vektoren zu erzeugen. Es gilt als eines der leistungsfähigsten frei verfügbaren Embedding-Modelle.

```
import { RunnableLambda } from '@langchain/core/runnables';
import { OllamaEmbeddings } from '@langchain/ollama';

const model = 'nomic-embed-text';
const baseUrl = 'http://localhost:11434';

const getEmbeddings = new RunnableLambda({
  func: async (text: string) => {
    const embeddings = new OllamaEmbeddings({
      model,
      baseUrl,
    });
    const embeddedDocuments = await embeddings.embedDocuments([text]);
    return embeddedDocuments;
  },

});
```

LangChain muss nur wissen, wie Ollama zu erreichen ist und welches Modell verwendet wird. Die eigentliche Arbeit erledigt Ollama.

Das Listing auf Seite 33 zeigt, wie sich Embeddings innerhalb von LangChain erzeugen lassen. Dabei kommt Ollama ins Spiel. Die Open-Source-Anwendung ist ein generisches Framework für LLMs. Auf der Homepage stehen Installer für Windows, macOS und Linux bereit. Nach der Installation lässt sich ollama via Kommandozeile steuern. Mit dem Befehl `ollama pull nomic-embed-text` lädt es das benötigte Sprachmodell.

LangChain verpackt die Kommunikation mit Ollama in Wrapper. Die `OllamaEmbeddings`-Klasse etwa spricht im Listing eine lokale Ollama-Instanz an, um die Embeddings mit dem gewünschten Modell zu erzeugen. Als Informationen benötigt der Konstruktor lediglich den Namen des Modells und die Adresse, über die Ollama seine Kommunikation abwickelt. Standardmäßig nutzt Ollama den lokalen Port 11434.

Führt die Applikation diesen Codeblock aus, baut die `OllamaEmbeddings`-Klasse die Verbindung zu Ollama auf, lädt das Modell und erzeugt die Embeddings für den übergebenen Text. Da nomic-embed-text 768-dimensionale Vektoren erzeugt, erzeugt die Funktion ein Array mit 768 Zahlen.

Der Vorteil dieser modularen Implementierung ist, dass Sie das Embedding-Modell einfach gegen ein anderes austauschen können. Um ein anderes Modell zu testen, installieren Sie dieses in Ollama und passen den Wert der model-Variablen an. Der Befehl `ollama pull mxbai-embed-large` installiert beispielsweise das mxbai-embed-large-Modell in Ollama.

Vektor-Silos

Je nach Umfang des Dokuments oder der Dokumente sowie der verwendeten Werkzeuge kann die Aufbereitung von Texten für ein RAG-System zeitaufwendig sein. Damit die Anwendung nicht bei jeder Anfrage auf ein Neues die Dokumente einlesen, aufteilen und in Vektoren umwandeln muss, speichert das System die Vektoren in einer spezialisierten Datenbank. Stellt der Nutzer eine Abfrage an das RAG-System, sucht die Applikation mit einem sogenannten Retriever nach passenden Vektoren und lädt die Informationen aus der Datenbank.

Vektordatenbanken erfahren mit der wachsenden Popularität von KI-Applikationen viel Aufmerksamkeit. So hat jeder größere Cloudanbieter eine solche Datenbank im Angebot. Es gibt eine Reihe von spezialisierten Vektordatenbanken von kommerziellen Produkten wie Pinecone bis zu frei und quelloffen

```
jo@jo-2402 ~ % ollama
Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model from a Modelfile
  show        Show information for a model
  run         Run a model
  stop        Stop a running model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  ps          List running models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.
jo@jo-2402 ~ % ollama pull nomic-embed-text
pulling manifest
pulling 970aa74c0a90... 100% ▕████████▏ 274 MB
pulling c71d239df917... 100% ▕████████▏  11 KB
pulling ce4a164fc046... 100% ▕████████▏   17 B
pulling 31df23ea7daa... 100% ▕████████▏  420 B
verifying sha256 digest
writing manifest
success
jo@jo-2402 ~ %
```

Ollama lässt sich von der Kommandozeile aus steuern.

```
import { Milvus } from '@langchain/community/vectorstores/milvus';
import { RunnableLambda } from '@langchain/core/runnables';
import { OllamaEmbeddings } from '@langchain/ollama';

const embeddingsModel = 'nomic-embed-text';
const collectionName = 'rag_collection';
const milvusTextFieldMaxLength = 2_000;

const storeVectors = new RunnableLambda({
  func: async (data: { texts: string[]; metadata: object }) => {
    console.log('Storing vectors');
    const embeddings = new OllamaEmbeddings({
      model: embeddingsModel,
      baseUrl: 'http://localhost:11434',
    });
    await Milvus.fromTexts(data.texts, data.metadata, embeddings, {
      collectionName,
      textFieldMaxLength: milvusTextFieldMaxLength,
      url: 'localhost:19530',
    });
    console.log('Vectors stored');
  },

});
```

Dieses Listing kombiniert das Umwandeln mit der Speicherung in der Vektordatenbank.

verfügbaren Lösungen wie Milvus und Chroma. Je nach Anforderungen an die Anwendung kommt entweder eine gehostete Datenbank oder eine lokale Lösung infrage.

Das Beispiel nutzt eine lokale Milvus-Instanz als Vektordatenbank. Milvus lässt sich sowohl zu Fuß installieren als auch über ein Image in der Docker Registry als Container betreiben. Milvus ist weitverbreitet, sodass für die Kommunikation zwischen der RAG-Applikation und der Datenbank verschiedene Konnektoren zur Verfügung stehen. Die Beispiel-Applikation nutzt die LangChain-Community-Erweiterung für Milvus. Der Code aus dem Listing oben kombiniert das Erzeugen der Embeddings und das Speichern in der Datenbank.

Der Milvus-Konnector von LangChain stellt die `fromTexts`-Methode zur Verfügung. Sie akzeptiert ein Array von Textblöcken, Metadaten, die den Text näher beschreiben, eine Embeddings-Instanz sowie ein Konfigurationsobjekt. Die Textblöcke kommen direkt vom Chunking. Die Metadaten können beispielsweise das eingelesene PDF-Dokument näher beschreiben. In diesem Fall stammen die Daten vom Laden des PDFs und müssen beim Chunking des Dokuments weitergereicht werden.

Das Embeddings-Modell ist das gleiche wie im vorangegangenen Schritt, nur dass der Milvus-Connector den Aufruf der `embedDocuments`-Methode selbst übernimmt. Das Konfigurationsobjekt erhält die `fromTexts`-Methode von Milvus. Es spezifiziert, in welcher Collection die Vektoren gespeichert werden und wie groß das Textfeld sein soll, in dem die Datenbank den Text ablegt, der zum Vektor gehört. Die `url`-Eigenschaft gibt an, über welche Adresse der Milvus-Server erreichbar ist.

Dieser letzte Baustein vervollständigt die Vorbereitung des RAG-Systems. Im nächsten Schritt fügt eine LangChain-Applikation alle Elemente zu einer

```
import { RunnableSequence } from '@langchain/core/runnables';

const sequence = RunnableSequence.from([loadPDF, splitText, storeVectors]);

await sequence.invoke(inputFile);
```

Die `RunnableSequence`-Klasse verkettet mehrere LangChain-Bausteine.

lauffähigen Kette zusammen, sodass die einzelnen Elemente funktionieren und ineinandergreifen.

Fertig verkettet

Die einzelnen Funktionen für die Vorbereitung der RAG-Applikation – das Laden, das Chunking, das Erzeugen der Embeddings sowie das Speichern in der Vektordatenbank – sind in Instanzen der `RunnableLambda`-Klasse gekapselt. Diese lassen sich mit der `from`-Methode der `RunnableSequence`-Klasse von LangChain zu einer Kette zusammenfassen. Das Listing oben zeigt den zugehörigen Code.

Die Applikation führt die so erzeugte Kette mit der `invoke`-Methode aus. Sie akzeptiert eine Eingabe für die erste Funktion in der Kette. In diesem Fall ist das der Name der PDF-Datei, die eingelesen werden soll. Am Ende der Implementierung liegen die Vektoren in der Milvus-Datenbank bereit zur Abfrage. Auf dieser Basis kann das RAG-System mit den Informationen arbeiten.

Sobald die Vektoren in der Datenbank liegen, müssen nur noch zwei Schritte umgesetzt werden, um vom Sprachmodell auf einen Prompt eine hilfreiche Antwort zu erhalten. Die Applikation erhält zunächst vom User einen Prompt und nutzt die Ähnlichkeitssuche der Vektordatenbank, um passende Textstellen zu finden.

Im zweiten Schritt übernimmt die Anwendung den Prompt und stellt die in der Datenbank gefundenen Textabschnitte dem angeschlossenen LLM als Kontext zur Verfügung. Das LLM liefert dann die Antwort. Mithilfe von LangChain lassen sich diese Schritte wieder zu einer kompakten Kette zusammenfügen. Den Code hierfür enthält das folgende Listing.

Die erste Säule der Implementierung bildet der Retriever. Er basiert auf dem Milvus-Connector mit den OllamaEnbeddings. Wichtig ist hier, dass die Applikation die gleichen Embeddings verwendet wie beim Einlagern der Daten in die Datenbank. Die meiste Arbeit übernimmt hier LangChain, sodass der für den Retriever erforderliche Code sehr kompakt gehalten werden kann.

Beim Auslesen der Chunks aus der Datenbank gibt es eine Optimierungsmöglichkeit: Standardmäßig begrenzt der Retriever die Ergebnisse, die er aus der Datenbank liest. Die Anzahl lässt sich erhöhen, indem man der `asRetriever`-Methode die Anzahl der gewünschten Chunks übergibt. Je mehr Chunks dem LLM zur Verfügung stehen, desto mehr Hintergrundwissen zur Formulierung der Antwort steht zur Verfügung. Zu viel Kontext kann sich jedoch negativ auf die Laufzeit der Applikation auswirken. Damit ist die Chunk-Anzahl eine mögliche Stellschraube, um die Qualität und Performance zu verbessern.

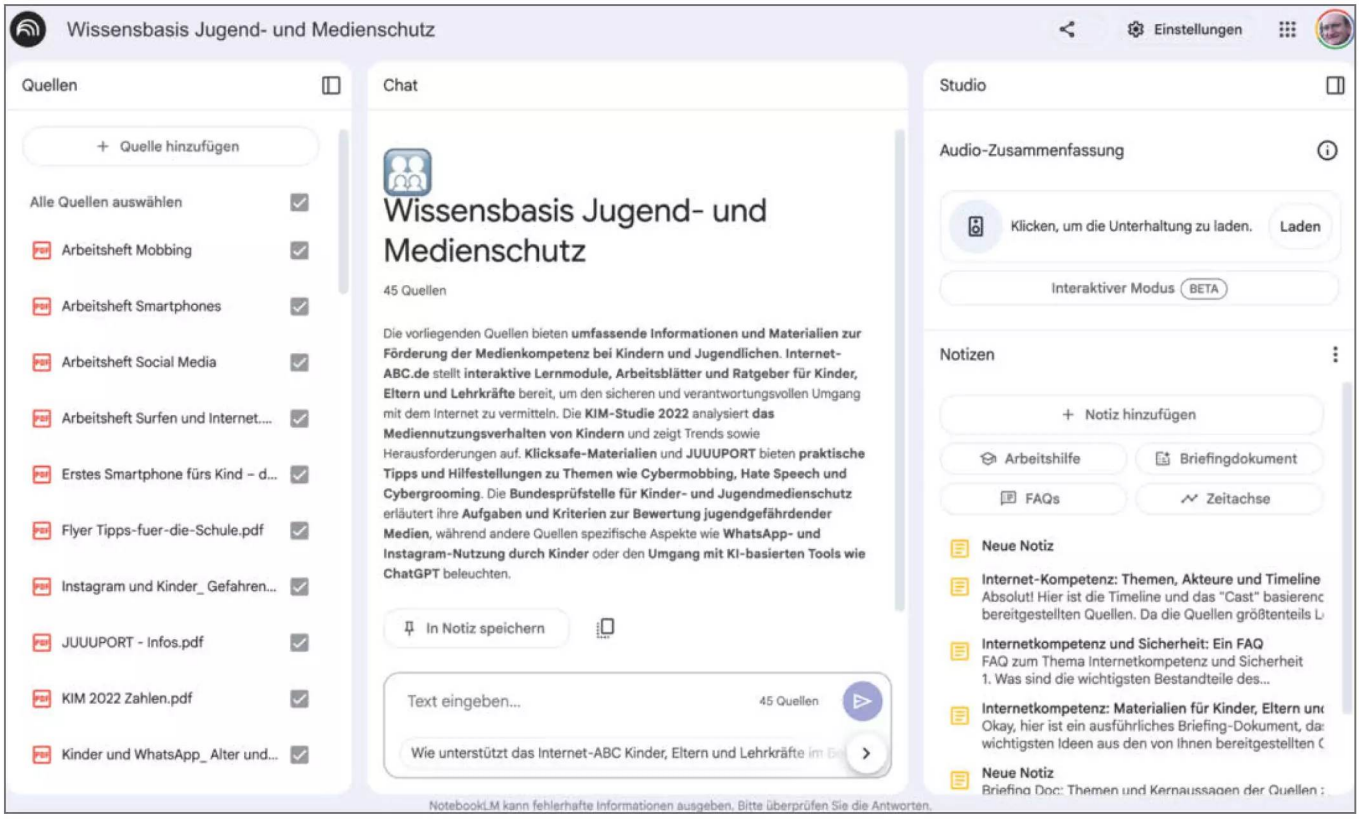
Die zweite Säule der Implementierung ist die Kommunikation mit dem LLM. Die Ausführungskette der LangChain-Applikation lädt im ersten Schritt mit dem Retriever die Daten aus der Datenbank und gibt die Anfrage des Users über eine `RunnablePassthrough`-Instanz an das nächste Element weiter. Hier setzt die Applikation den Kontext und den Prompt in ein Prompt-Template ein, um dem LLM klare Anweisungen zu geben.

Der nächste Schritt ist die Kommunikation mit dem LLM. Das Beispiel nutzt ein lokales LLama-3.2-Modell, das in Ollama ausgeführt wird. Auch in diesem Schritt gibt es kaum Einschränkungen, was die verfügbaren Modelle angeht. Die Varianten reichen von leichtgewichtigen lokalen Modellen bis hin zu leistungsstarken gehosteten Modellen wie die GPT-Modelle von OpenAI oder Claude von Anthropic.

Nachdem das Modell seine Antwort anhand der Eingaben der Applikation erzeugt hat, extrahiert der `StringOutputParser` die Antwort und gibt die Zeichen-

```
import { formatDocumentsAsString } from 'langchain/util/document';
import { PromptTemplate } from '@langchain/core/prompts';
import {
  RunnableSequence,
  RunnablePassthrough,
} from '@langchain/core/runnables';
import { StringOutputParser } from '@langchain/core/output_parsers';
import { Milvus } from '@langchain/community/vectorstores/milvus';
import { Ollama, OllamaEmbeddings } from '@langchain/ollama';

const collectionName = 'rag_collection';
const llm = 'llama3.2';
const embeddingsModel = 'nomic-embed-text';
const model = new Ollama({
  model: llm,
});
const embeddings = new OllamaEmbeddings({
  model: embeddingsModel,
  baseUrl: 'http://localhost:11434',
});
const vectorStore = new Milvus(embeddings, {
  collectionName,
  url: 'localhost:19530',
});
const retriever = vectorStore.asRetriever(10);
const prompt =
  PromptTemplate.fromTemplate(`Answer the question based only on the following context:

{context}

Question: {question}`);

const chain = RunnableSequence.from([
  {
    context: retriever.pipe(formatDocumentsAsString),
    question: new RunnablePassthrough(),
  },
  prompt,
  model,
  new StringOutputParser(),
]);

const result = await chain.invoke(
  'How to create a colored output on the console with Node.js?'
);

console.log(result);
```

Auch das Retrieval der passenden Chunks und der Bau des Prompts bilden eine Kette.



Googles NotebookLM ist eine für jedermann zugängliche RAG-Anwendung, die bis zu 50 Dokumente erschließt.

kette zurück. Diese kann entweder, wie im Beispiel, direkt auf der Konsole ausgegeben oder an einen Client oder ein anderes System gesendet werden.

Weiterschrauben

Mit Retrieval Augmented Generation können Sprachmodelle auf aktuelle und vertrauliche Daten zugreifen, ohne dass diese öffentlich verfügbar sein müssen oder ein Modell extra damit trainiert werden muss. Das eröffnet neue Möglichkeiten für LLMs. So kann man einen Chatbot bauen, der Fragen zu aktuellen Produkten beantwortet.

Frameworks wie LangChain helfen dabei, den modularen RAG-Prozess abzubilden. Der größte Aufwand liegt bei einer RAG-Applikation in der Vorbereitung der Daten. Ist die Grundlage schlecht, können keine guten Resultate entstehen. Dabei gibt es zahlreiche Regler, die sich sowohl auf die Performance des Prozesses als auch auf die Qualität der Ergebnisse auswirken. Dazu zählen beispielsweise die Chunking-Strategie und die Auswahl des Embeddings-Modells. Beim Information Retrieval und der Textgenerierung gibt es ebenfalls zahlreiche Konfigurationsmöglichkeiten wie die Anzahl der ausgelesenen Chunks und die Formulierung des Prompt-Templates.

So gibt es nicht die eine richtige Lösung für ein RAG-System. Stattdessen kann man erst mal loslegen und sich von einer ersten Implementierung mit viel Feintuning an ein optimales Ergebnis herantasten. Unter dem c't-Link finden Sie ein fertiges, einfaches Projekt, das die in diesem Artikel beschriebenen Schritte umsetzt. (jo) c't

Komplettes Projekt:
ct.de/w2ch

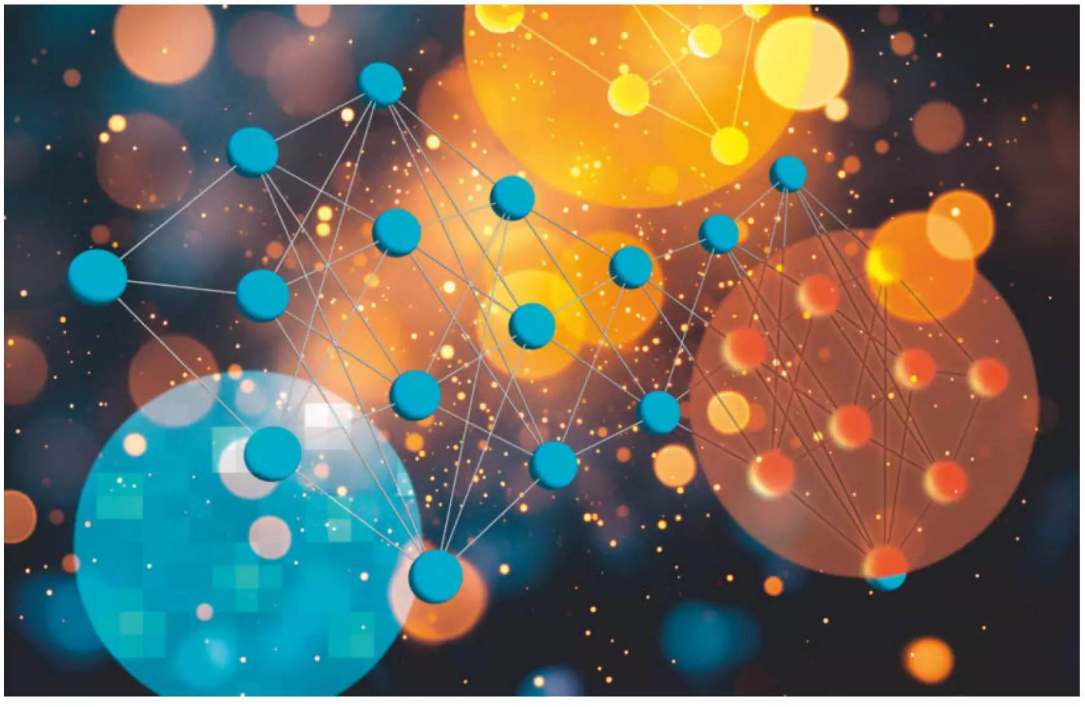


Bild: KI, Collage c't

KIs ohne High-End-Hardware betreiben

Theoretisch kann jeder das 671-Milliarden-Parameter-Modell DeepSeek herunterladen und selbst betreiben. Theoretisch. Praktisch kann sich kaum jemand die teuren Grafikkarten dafür leisten. Deshalb gibt es komprimierte Versionen davon. Wir erklären, was sich dahinter verbirgt.

Von **Andrea Trinkwalder**

DeepSeek gilt als besonders effizientes Large Language Model (LLM), das nicht nur ressourcenschonend trainiert wurde, sondern dem Spitzenreiter von OpenAI in bestimmten Disziplinen ebenbürtig oder überlegen ist. Weil sogar die größte Modellversion als Open-Weights-Modell zum Download steht, kann es theoretisch jeder selbst auf seinen Servern betreiben. Das ist deshalb interessant, weil man damit den Zugriff via Web-Interface und Mobil-App umgeht, der recht strengen chinesischen Inhaltsfiltern unterliegt und vor allem für Firmen datenschutzrechtlich heikel ist.

Mittlerweile kursieren zahlreiche DeepSeek-Ableger, die deutlich weniger Ressourcen benötigen als das große Basismodell – und teils verwirrende Namen tragen. Sie entstanden mithilfe unterschiedlicher Komprimierungstechniken aus der recht unhandlichen und ressourcenintensiven Ursprungsversion.

Bei näherem Hinsehen stellt man fest: Viele dieser vermeintlichen DeepSeek-Varianten sind eigentlich Llama- oder Qwen-Architekturen, denen quasi ein Teil des DeepSeek-Gehirns transplantiert wurde. Wir geben einen Überblick über die gängigen Verfahren und wie sie sich auswirken.

Bedingt betriebsbereit

Auch wenn Nvidia-Anleger wegen DeepSeek panisch ihre Aktien abgestoßen haben: Daraus sollte man nicht den falschen Schluss ziehen, dass das chinesische Open-Weights-Sprachmodell einfach so in einem x-beliebigen Firmenrechenzentrum oder gar auf jedem Notebook läuft. Das ausgewachsene 671-Milliarden-Parameter-Reasoning-Modell DeepSeek R1 benötigt 720 GByte RAM. Anders ausgedrückt: Es passt mit Müh und Not auf ein 200.000 US-Dollar teures DGX-H100-System von Nvidia mit acht H100-GPUs. So etwas stellt man sich nicht mal eben ins Rechenzentrum, aber es ist immer noch um einige Größenordnungen günstiger als die Hardware, auf der etwa OpenAI sein GPT betreibt. OpenAI und Microsoft wollen die Server-Infrastruktur in den nächsten Jahren für Hunderte Milliarden US-Dollar ausbauen.

Damit Sprachmodelle dennoch einigermaßen wirtschaftlich oder gar lokal auf schwächerer Hardware betrieben werden können, haben Machine-Learning-Spezialisten diverse sogenannte Kompressionsmethoden entwickelt: Quantisierung, Pruning und Wissensdestillation. Sie sollen unnötige Rechenoperationen einsparen, aufwendige vereinfachen oder die Modelle auf ein optimales Maß verkleinern, und zwar bei möglichst gleichbleibender Ausgabequalität.

Auch von DeepSeek gibt es quantisierte und destillierte Versionen. Einige hat der Hersteller selbst veröffentlicht, andere stammen von externen Entwicklern. Die beherbergen aber in einigen Fällen nicht das echte DeepSeek, sondern manchmal lediglich ein DeepSeek-imitierendes Llama oder Qwen. Die dafür verwendete Technik erklären wir im Abschnitt „Destillierung: Hochprozentiges Wissen“ weiter unten.

Quantisierung: Genauigkeit verringern

Um zu verstehen, wie Machine-Learning-Modelle eingedampft werden, hilft ein kurzer Überblick über die Struktur und die Funktionsweise tiefer neuronaler Netze, die auch die Grundbausteine der großen Sprachmodelle sind. Sie bestehen aus mehreren Schichten von künstlichen Neuronen, die über Verknüpfungen, sogenannte Kanten, miteinander verbunden sind. Die Stärke jeder Verbindung wird durch ihr sogenanntes Kantengewicht w ausgedrückt und der Wert x jedes Neurons berechnet sich als Summe der Reize, die über die Kanten eintreffen. Liegen diese über der individuellen Reizschwelle des Neurons, feuert es. Die Gewichte bezeichnet man auch als lernbare Parameter: Sie werden während des Trainings anhand unzähliger Beispiele (Token) so lange justiert, bis das Netz die zugrunde liegende Aufgabe bewältigt.

LLMs lernen zum Beispiel anhand von Lückentexten, Sätze Wort für Wort zu vervollständigen – am Ende sind sie in der Lage, selbst Geschichten oder längere Texte in unterschiedlichem Stil zu verfassen. Gesteuert wird dieses permanente Aktualisieren über eine Fehlerfunktion. Sie vergleicht die Vorhersage (also das vom Netz berechnete Ergebnis) mit der Musterlösung (Ground Truth) und passt anhand der Ausprägung der Abweichung die hauptsächlich dafür verantwortlichen Parameter stärker an als diejenigen, die weniger Anteil am Fehler hatten. Mathematisch ausgedrückt: Dazu muss in einem Näherungsverfahren das Minimum dieser Fehlerfunktion gefunden werden, und dieses Näherungsverfahren heißt Gradientenabstieg. Details dazu siehe [1]; hier festzuhalten bleibt, dass im Zuge dessen sehr viele partielle Ableitungen (Gradienten) zu berechnen sind.

Trickreiches Training

Weil die größten Sprachmodelle wie GPT oder Gemini Hunderte Milliarden Parameter haben und mit Milliarden von Token trainiert werden, kommen während eines solchen Trainings eine Menge einzelne Rechnungen (Matrixoperationen) zustande: In jedem Durchlauf müssen die Gradienten berechnet und die Gewichte nachjustiert werden. Besonders viel Rechen- und Speicherkapazität benötigen diese mathematischen Operationen, wenn sie mit der höchstmöglichen Präzision ausgeführt werden, also mit Gleitkommazahlen in 32 Bit (Floating Point, FP32). Das ist aber nicht durchgängig notwendig.

Quantisierung

Bei der Quantisierung bleibt die Struktur des künstlichen neuronalen Netzes unangetastet. Der Rechenaufwand wird reduziert, indem man bestimmte Rechenschritte mit geringerer Genauigkeit ausführt. Dazu werden die Gewichte bei Bedarf von FP32 in FP16, FP8 o. ä. konvertiert.

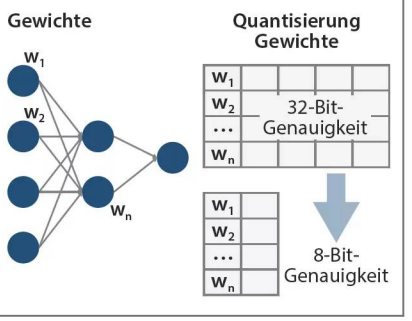


Bild: c't, Quelle: Springer Nature

Insbesondere im praktischen Einsatz des fertig trainierten Sprachmodells (Inferenz), aber auch schon während der Trainingsphase lässt sich der Rechenaufwand erheblich reduzieren, wenn man ein paar Nachkommastellen einspart, also sich etwa mit 16-Bit-Gleitkommazahlen (FP16) oder noch geringerer Genauigkeit begnügt. Quantisierung bedeutet also: Rechnen mit reduzierter Genauigkeit.

Während des Trainings ist allerdings Fingerspitzengefühl gefragt, denn wer flächendeckend mit halber Genauigkeit rechnet, macht das Modell instabil. Das liegt daran, dass die korrekte Prognose manchmal nur knapp vor der oder den falschen möglichen Lösungen liegt, etwa wenn ein Satz doppeldeutig ist oder der Kontext sich erst am Ende vage andeutet. Solche feinen Nuancen dürfen während des Trainings nicht verlorengehen, sonst kippen Entscheidungen an neuralgischen Punkten zu oft in die falsche Richtung.

Deshalb hat sich der sogenannte Mixed-Precision-Ansatz etabliert, bei dem man die Gewichte immer in voller Genauigkeit vorhält, aber die Berechnung der Gradienten in halber Genauigkeit stattfindet: Dazu werden zunächst die FP32-Gewichte in FP16 konvertiert, um die Gradienten effizient berechnen zu können, sprich: Rechenzeit und Speicher zu sparen. Diese FP16-Gradienten werden anschließend in FP32 umgewandelt, um wiederum die Original-FP32-Gewichte zu aktualisieren. Dieser Schritt muss in voller Genauigkeit ausgeführt werden, um numerische Stabilität zu gewährleisten. Ansonsten kann es passieren, dass Gradienten während des Optimierungsprozesses „verschwinden“ oder „explodieren“: Im ersten Fall bricht das Training zu früh ab, im zweiten werden sie so groß, dass es nie zum Ende kommt.

Die DeepSeek-Entwickler haben ihr Basismodell V3, das dem neuen Reasoning-Modell R1 zugrunde liegt, sogar zu einem Großteil mit FP8-Genauigkeit trainiert. Wie sie im zugehörigen Aufsatz erläutern, führen sie die meisten Matrixmultiplikationen (General Matrix Multiplications, GEMM) in FP8-Genauigkeit aus, während Operatoren, die zu sensibel auf verringerte Präzision reagieren, in höherer Genauigkeit bleiben; alle Quellen siehe ct.de/w2wq.

Das ist bemerkenswert, aber wer mit einem selbst aufgesetzten LLM liebäugelt, dürfte sich eher für die Quantisierung im laufenden Betrieb interessieren. Hier kann man sogar auf 4 Bit oder noch weniger heruntergehen.

Ein Durchbruch insbesondere für den Betrieb kleiner bis mittelgroßer Netze war die im Frühjahr 2024 vorgestellte 4-Bit-Quantisierung AWQ (Activation-Aware Weight Quantization). Sie reduziert das Gros der Modellgewichte auf 4 Bit und belässt einen geringen, aber für die Vorhersagequalität entscheidenden Teil in FP16-Genauigkeit: diejenigen 0,1 bis 1 Prozent der Gewichte, die besonders häufig aktiviert werden. Damit einher gingen erhebliche Performancesteigerungen, die allerdings mit Optimierungen für Nvidias CUDA-Befehlssätze erzielt wurden. Das bedeutet: Damit AWQ nicht nur Speicher spart, sondern auch maximal schnell rechnet, muss die gesamte Pipeline auf Nvidia-GPUs ausgerichtet sein. Details und Benchmarks zu diversen Quantisierungsverfahren finden Sie in [2].

Flexibler ist dynamische Quantisierung per GGUF (Georgi Gerganov Unified Format), weil es auch die Fähigkeiten moderner CPUs ausnutzt: etwa Intels neue Befehlssatzerweiterungen Advanced Matrix Extensions (AMX), die Matrixmultiplikationen beschleunigen.

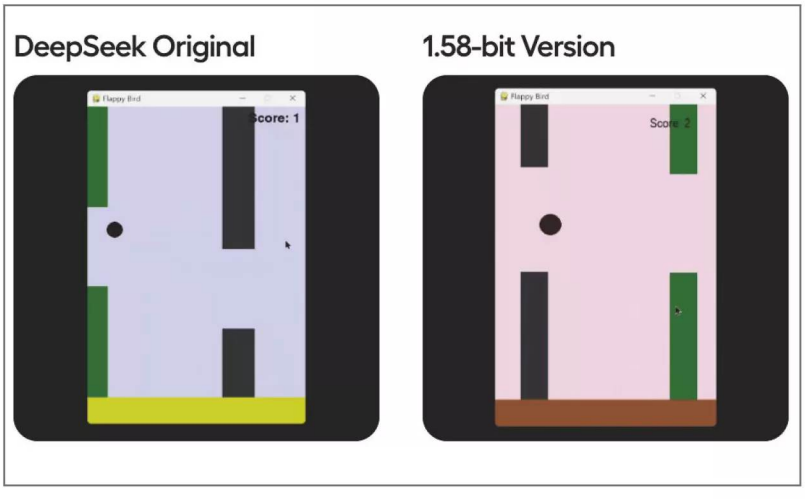
Auf Basis von ebendiesem GGUF haben die Finetuning-Experten von Unsloth AI die derzeit am stärksten komprimierte 671B-Variante von DeepSeek R1 veröffentlicht: eine 1,58-Bit-Quantisierung, die nur 131 GByte und damit 20 Prozent des Originals belegt.

Die Finetuning-Schmiede Unsloth AI hat eine dynamische Quantisierung für DeepSeek ausgetüftelt und stellt vier Varianten von 1,58 bis 2,51 Bit zum Download.

1. Dynamic Quantized versions

We provide 4 dynamic quantized versions. The first 3 uses an importance matrix to calibrate the quantization process (imatrix via llama.cpp) to allow lower bit representations. The last 212GB version is a general 2bit quant with no calibration done.

MoE Bits	Disk Size	Type	Quality	Link	Down_proj
1.58-bit	**131GB**	IQ1_S	Fair	Link	2.06/1.56bit
1.73-bit	**158GB**	IQ1_M	Good	Link	2.06bit
2.22-bit	**183GB**	IQ2_XXS	Better	Link	2.5/2.06bit
2.51-bit	**212GB**	Q2_K_XL	Best	Link	3.5/2.5bit



Dass die 1,58-Bit-Variante geistig durchaus auf der Höhe bleibt, demonstriert Unsloth mit einem eigenen Benchmark: der Aufgabe, ein Flappy-Bird-Spiel zu programmieren.

Dazu haben die Entwickler die Modellarchitektur analysiert und bestimmte Layer identifiziert, die in 4 Bit oder höherer Genauigkeit bleiben mussten, während sie die meisten Layer innerhalb der Teilnetze (MoE-Layer), aus denen DeepSeek besteht, auf 1,5 Bit setzen konnten. DeepSeek ist als Zusammenschluss von Experten-Netzwerken (Mixture of Experts, MoE) konstruiert, sodass während Training und Inference immer nur diejenigen Teile aktiv werden müssen, die für die Lösung der konkreten Aufgabe zielführend sind.

Auf ihrer Website schreiben die Entwickler, dass man ihr 1,58-Bit-Modell mit nur 20 GByte RAM betreiben kann, wenn auch sehr langsam.

Pruning: Baumschnitt für KNNs

Die Quantisierung reduziert lediglich die Genauigkeit der Werte in den Matrizen, verringert aber die Größe oder Anzahl der Matrizen nicht zwangsläufig, sprich: Die Netzarchitektur bleibt unverändert und dicht verzweigt, die Anzahl der Rechenoperationen bleibt

Pruning

Pruning-Techniken dünnen das künstliche neuronale Netz aus, indem gezielt Neuronen oder ganze Schichten entfernt werden, die sich nicht wesentlich auf die Vorhersagegenauigkeit auswirken. Dadurch reduziert sich die Anzahl der Rechenoperationen.

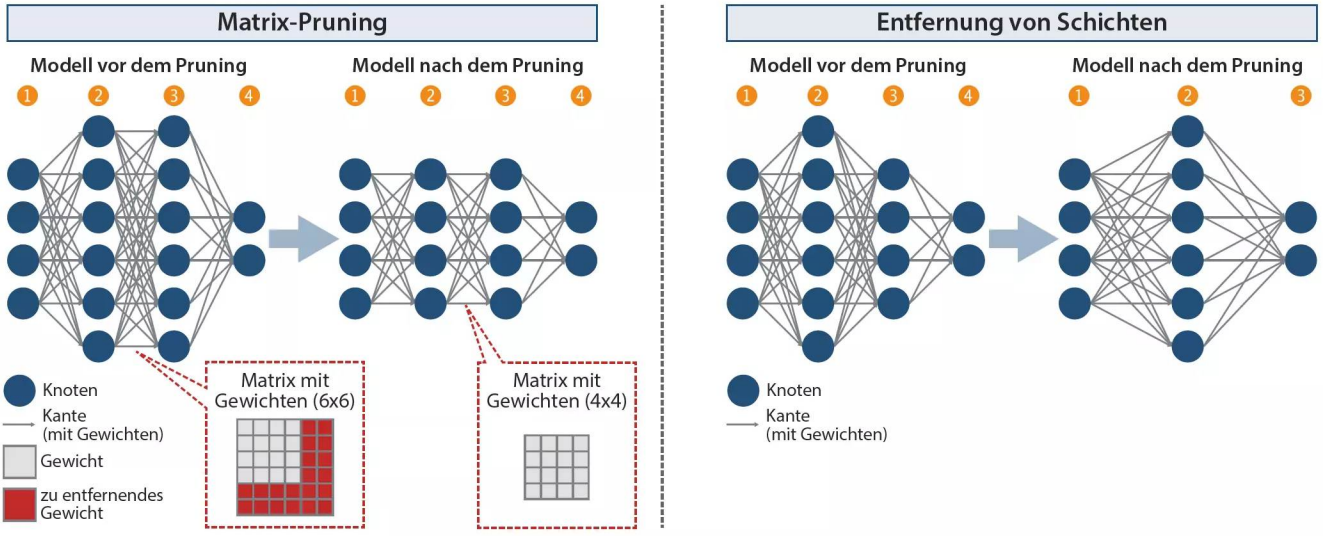


Bild: c't, Quelle: Springer Nature

Wissensdestillation

Bei der Wissensdestillation versucht man, die Art und Weise, wie ein sehr großes KI-Modell mit hunderten Milliarden Parametern Vorhersagen trifft (Lehrermodell) auf ein kleineres mit unter hundert oder unter zehn Milliarden zu übertragen (Schülermodell). Dazu wird das kleinere Netz nicht nur mit den Trainingsdaten und den zugehörigen Labels (Kategorien, Hard Labels) gefüttert, sondern auch mit der Ausgabe – also den Vorhersagen – des Lehrermodells (Soft Labels).

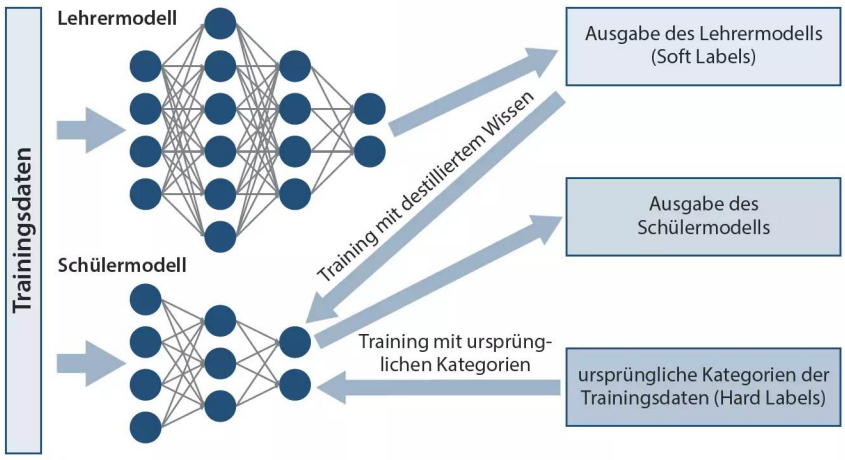


Bild: c't, Quelle: Springer Nature

Literatur

[1] Andrea Trinkwalder, Die Mathematik hinter neuronalen Netzen, c't 11/2024, S. 22

[2] René Peinl, Komprimierte KI, Wie Quantisierung große Sprachmodelle verkleinert, c't 2/2025, S. 120

Publikationen und Modelle:

ct.de/w2wq

gleich. Ähnlich wie man beim professionellen Baumschnitt nur die stärksten und für einen gleichmäßigen Wuchs wichtigen Äste behält, lassen sich aber auch künstliche neuronale Netze so geschickt ausdünnen, dass deren Prognosequalität nur geringfügig leidet: entweder, indem man nur einzelne Gewichte entfernt (unstrukturiertes Pruning) oder ganze Neuronen beziehungsweise Filter (strukturiertes Pruning).

Pruning ist ein Prozess aus Training, Ausdünnen und Finetuning, der manchmal auch iterativ abläuft. Dabei wird das KNN zunächst trainiert und anschließend analysiert, welche Gewichte, Neuronen oder auch ganze Filter verzichtbar sind, weil sie nur schwach oder selten aktiviert wurden. Darauf folgt ein Finetuning des ausgedünnten Konstrukts, um den Verlust an Komplexität zu kompensieren. Häufig fungiert das größere Netz dabei als Trainer, damit möglichst viele seiner höheren Fähigkeiten auch in dem kleineren Ableger erhalten bleiben.

So entstehen mit vergleichsweise geringem Aufwand kompaktere Versionen eines großen Modells – eine echte Alternative zur klassischen Herangehensweise vieler größerer KI-Schmieden. Meta hat beispielsweise noch alle drei Varianten (8B, 70B und 405B) seines Open-Weights-Modells Llama von Grund auf trainiert.

Destillierung: Hochprozentiges Wissen

Während die Quantisierung auf numerischer und das Pruning auf struktureller Ebene stattfindet, ist das Destillieren (Model Distillation, MD) inhaltlicher Natur: Es ist die Kunst, die Essenz des in einem Hunderte Milliarden Parameter großen Basismodell gespeicherten Wissens auf ein kleineres Modell zu übertragen. Ziel dabei ist, dass dieses annähernd so robust und genau arbeitet wie das große Vorbild. Der Wissenstransfer gelingt mithilfe eines Lehrer-Schüler-Konstrukts, in dem das kleine Modell lernt, die Antworten beziehungsweise das Antwortverhalten des großen nachzuahmen. Es handelt sich also um eine spezielle Trainingsstrategie, während Quantisierung und Pruning überwiegend auf fertig trainierte Modelle angewendet werden.

Mithilfe der Wissensdestillation können LLM-Entwickler kleinere, ressourcensparende Versionen ihrer riesigen Basismodelle wie GPT oder Gemini erzeugen. Aber nicht nur das: Praktischerweise funktioniert das auch von Hersteller zu Hersteller, also etwa von GPT-4 auf Llama und DeepSeek oder von DeepSeek auf Llama und Qwen. Welchen Anteil die Wissensdestillation tatsächlich am Training von Llama oder DeepSeek hat, ist übrigens genauso wenig dokumentiert wie der Anteil von urheberrechtlich geschütztem Material in den ChatGPT-Trainingsdaten. OpenAI sind diese Praktiken jedenfalls ein Dorn im Auge, weshalb es in seinen Nutzungsbedingungen verbietet, ChatGPT & Co. für solche Zwecke einzuspannen.

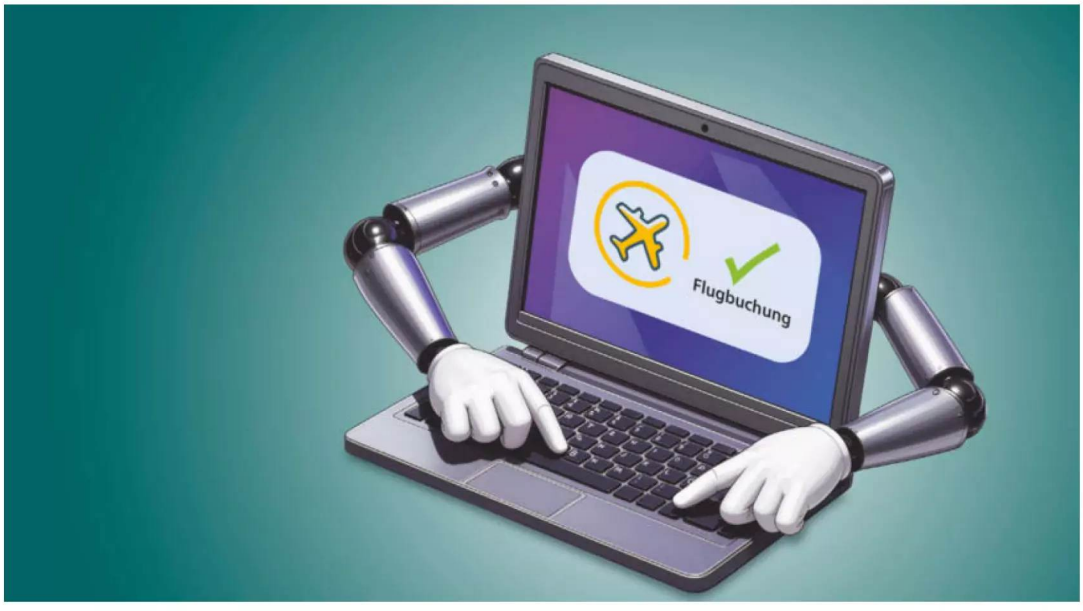
Im einfachsten Fall werden beide Netze mit denselben Trainingsdaten gefüttert, also im Fall von Sprachmodellen mit Satzfragmenten, die sie vervollständigen sollen. Das Lehrermodell gibt nun eine Prognose darüber ab, welche Token den Satz korrekt fortsetzen könnten, also eine Wahrscheinlichkeitsverteilung. Das kleinere Schülermodell soll nun lernen, zu jedem Trainings-Sample die gesamte, vom Lehrer vorgegebene Verteilung zu reproduzieren, anstatt nur das passendste Token auszuwählen.

Bei einem solchen rein ausgabeorientierten Training gehen allerdings wertvolle Verhaltensmuster des großen Vorbilds verloren, weil sie wichtige Prozesse, die innerhalb der komplexen Sprach- oder multimodalen Modelle ablaufen, zu wenig berücksichtigen. Neuere, speziell für Transformer-Netze entwickelte Verfahren wie etwa das von Nvidia vorgestellte Minitron beziehen deshalb auch den Aufmerksamkeitsmechanismus mit ein, der ja primär die „Denkweise“ solcher Netzarchitekturen steuert. Die derzeit von DeepSeek verfügbaren destillierten Varianten sind übrigens Llama- oder Qwen-Architekturen, die vom 671B-DeepSeek geschult wurden.

Wem es nützt

Von Kompressionstechniken und insbesondere der Wissensdestillation profitierten vor allem Start-ups und Forschungseinrichtungen, die weder die finanziellen noch die technischen oder personellen Mittel für das Training eines Hunderte-Milliarden-Parameter-LLM aufbringen können. Aber auch lokal beziehungsweise auf Mobilgeräten laufende sowie auf bestimmte Anwendungen spezialisierte Sprachmodelle rücken in greifbare Nähe.

Allerdings lassen sich die höheren Fähigkeiten eines 671-Milliarden-Parameter-Netzes wie DeepSeek nicht beliebig konservieren und auf das Sieben-Milliarden-Geflecht eines anderen Herstellers transplantieren. Wer also selbst mit dem Betrieb eines eigenen DeepSeek liebäugelt, muss sich ein wenig mit den verschiedenen Varianten auseinandersetzen: Hinter destillierten DeepSeeks verbirgt sich Stand heute meist ein Llama oder Qwen. (atr) c't



(Bild: Martina Bruns/KI/heise medien)

MCP: KI greift auf Apps und Daten zu

Mit dem Model Context Protocol können Sie KI-Sprachmodelle im Nu mit Apps und Daten verzahnen. Schon heute können ChatGPT & Co. selbstständig tausende Apps steuern, darunter Browser, Business-Apps, Spotify und Ihr Smart Home. Das eröffnet unzählige neue Einsatzmöglichkeiten, birgt aber auch neue Risiken. Ein Überblick.

Von **Jo Bager und Ronald Eikenberg**

KI-Sprachmodelle werden zu Agenten. Sie treffen selbstständig Entscheidungen, interagieren mit Daten und steuern Werkzeuge, um Aufgaben zu erledigen. Bei Googles Entwicklerkonferenz I/O waren diese KI-Agenten eines der Hauptthemen. Der Konzern will zum Beispiel seine Suchmaschine mit agentischer KI aufpeppen (siehe c't 13/2025, S. 40).

Windows-Nutzer werden KI-Agenten demnächst in ihrem Betriebssystem vorfinden, denn Microsoft baut solche Helfer in Windows ein. Statt sich selbst durch Anwendungen zu klicken, soll der Anwender künftig seinem PC einfach sagen, was der machen soll. Der PC erledigt die Aufgaben dann, indem er installierte Anwendungen steuert und auf das Dateisystem und Webdienste zugreift (siehe c't 13/2025, S. 32).

Eine Schlüsseltechnik, die das ermöglicht, ist das Model Context Protocol (MCP). Dieser Artikel erklärt, was sich dahinter verbirgt und warum MCP die An-

wendungsmöglichkeiten für Sprachmodelle erheblich erweitert. Der Artikel auf Seite 52 zeigt Beispiele für den Einsatz von MCP – und die Probleme, mit denen man dabei rechnen muss. Im Beitrag auf Seite 58 demonstrieren wir Ihnen, wie Sie einen eigenen MCP-Server programmieren. Dabei kann hinsichtlich Security einiges schiefgehen: Im Artikel auf Seite 66 diskutieren wir die durchaus vorhandenen Sicherheitsprobleme ausführlich.

Kontext ist alles

Um den Nutzen des Model Context Protocol nachvollziehen zu können, ist es sinnvoll, sich die rasante Evolution von Sprachmodellen vor Augen zu führen. Sprachmodelle sind in ihrer ursprünglichen Form nicht mehr als Textgeneratoren, die auf das gelernte Wissen beschränkt sind. Sie können zwar komplexe Texte generieren und analysieren, haben aber keinen Zugang zu aktuellen Informationen und externen Datenquellen. Bei spezifischen, aktuellen oder kontextabhängigen Informationen müssen sie passen oder halluzinieren schon mal. Externe Dienste oder Anwendungen steuern können sie erst recht nicht.

In den vergangenen Jahren hat man diese Textmaschinen um immer neue Fähigkeiten erweitert. Retrieval-Augmented Generation (RAG) etwa verbindet Sprachmodelle mit externen Wissensdatenbanken (siehe auch auf Seite 26). Dabei werden bei einer Anfrage an ein Sprachmodell zunächst relevante Dokumente oder Textpassagen abgerufen und dem Sprachmodell als Kontext bereitgestellt. So kann es auf aktuelle oder domänenspezifische Informationen zugreifen, ohne dafür neu trainiert werden zu müssen.

Function Calling geht einen Schritt weiter und ermöglicht Sprachmodellen, selbstständig externe Apps und Dienste zu steuern. Dabei stellt man dem Modell eine Liste verfügbarer Funktionen der jeweiligen Anwendung bereit, die gesteuert werden soll, inklusive deren Parameter und Beschreibungen. Das ist jederzeit möglich, ein erneutes Training ist dazu nicht nötig. Function Calling transformiert Sprachmodelle von reinen Textjongleuren zu aktiven Agenten, die mit der digitalen Umgebung interagieren.

Für das Function Calling unterhält allerdings jeder Sprachmodellanbieter (OpenAI, Anthropic, Google etc.) eigene Programmierschnittstellen und Implementierungsansätze. Das hat zu einer fragmentierten Landschaft geführt: Entwickler müssen für jedes Sprachmodell eigene Funktionsdefinitionen und Abrufmechanismen implementieren.

USB-C für Sprachmodelle

MCP ist die Antwort auf dieses Problem: Es bietet ein einheitliches Protokoll, das sprachmodell- und toolübergreifend funktioniert. Entwickler müssen die Schnittstelle zu eigenen Anwendungen – seien es Datenbanken, APIs oder lokale Tools – nur einmal als sogenannte MCP-Server implementieren. Mit dem können dann verschiedene KI-Modelle interagieren. Dies reduziert den Entwicklungsaufwand erheblich. MCP wird daher gerne „USB-C-Standard“ für KI-Tools bezeichnet – einmal implementiert, überall nutzbar.

Anthropic hat das Model Context Protocol erst im November 2024 vorgestellt. Die Welt schien auf so etwas wie MCP gewartet zu haben, denn es hat sich schnell als Quasistandard in der Branche durchgesetzt. OpenAI hat ebenso angekündigt, MCP zu unterstützen wie kürzlich Google und Microsoft. Es gibt bereits Tausende MCP-Server für die verschiedensten Anwendungen und Dienste.

So funktioniert MCP

MCP basiert auf einer Client-Server-Architektur, die bewusst einfach gehalten ist. Dabei gibt es drei Hauptakteure. Der MCP-Host ist die Schlüsselkomponente, die MCP-Funktionen in ein größeres System integriert. Der Host verwaltet die Verbindungen zu den MCP-Servern und stellt ihre Funktionen dem Sprachmodell zur Verfügung.

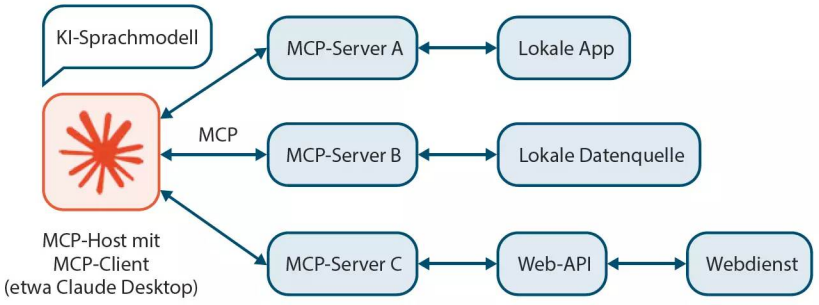
Die Desktop-Anwendung von Claude ist ein Beispiel für einen MCP-Host, aber auch Programmierumgebungen und Anwendungen für den Unternehmenseinsatz können MCP-Hosts sein. Der MCP-Client ist die technische Implementierung innerhalb des Hosts, die die Low-Level-Kommunikation mit den MCP-Servern abwickelt.

MCP-Server sind die Adapter zwischen KI-Systemen und konkreten Apps, Datenquellen oder APIs. Ein MCP-Server kann zum Beispiel eine Datenbank abfragen, ein anderer versendet E-Mails, ein dritter greift auf Kalenderdaten zu. Die KI kann all diese Server in Kombination nutzen, um komplexe Aufgaben zu erledigen.

Server und Client senden sich JSON-Schnipsel im Format JSON-RPC 2.0. Wie die im Detail aussehen, lesen Sie im Artikel auf Seite 58. Ein zentrales Konzept von MCP sind „Capabilities“. Das sind klar definierte Fähigkeiten, die ein Server bereitstellt. Jeder MCP-Server kann drei Arten dieser Capabilities bereitstellen: Tools sind Aktionen, die ausgeführt werden

Model Context Protocol: So steuern KI-Agenten Apps und Dienste

Das Model Context Protocol (MCP) spezifiziert einen einheitlichen Weg, wie KI-Sprachmodelle (Large Language Model, LLM) auf lokale Apps und Daten sowie Webdienste zugreifen können. Der MCP-Client (meist in den MCP-Host integriert) bringt über die MCP-Server in Erfahrung, welche Funktionen zur Verfügung stehen und bietet sie dem Sprachmodell an. Möchte das Sprachmodell eine der Funktionen verwenden, wird sie auf dem MCP-Server aufgerufen. Dieser kümmert sich darum, die App oder den Webdienst passend anzusprechen, oder er liest die angeforderten Daten direkt ein. Anschließend gibt es eine Rückmeldung über den MCP-Host an das Sprachmodell, das daraufhin weitere Schritte planen kann.



können, etwa „E-Mail senden“ oder „Datei erzeugen“. Tools können etwas in der echten Welt verändern (in einer Datenbank zum Beispiel oder im Smart Home), oder etwas berechnen.

Resources sind in der MCP-Terminologie Inhalte, die gelesen werden können, beispielsweise Dokumente, Datenbankeinträge oder API-Antworten. Im Unterschied zu Tools darf ihr Aufruf nichts verändern. Prompts schließlich sind vorgefertigte Textbausteine, die als Vorlagen für häufige Anfragen dienen. Ein GitHub-Server könnte beispielsweise Tools zum Anlegen von Issues anbieten, Resources für das Lesen von Repositories und Prompts für typische Entwickler-Workflows.

Sobald ein MCP-Client eine Verbindung zu einem MCP-Server aufbaut, fragt er ab, welche Funktionen dieser anbietet. Der Server antwortet mit einer detaillierten Liste seiner Capabilities, einschließlich ihrer Beschreibungen und benötigter Parameter. Diese Selbstauskunft macht das System flexibel: Neue Server können hinzugefügt werden, ohne dass bestehende Clients angepasst werden müssen, da sie automatisch beim Verbindungsaufbau lernen, mit den neuen Funktionen umzugehen.

Die JSON-Nachrichten können auf zwei Transportwegen verschickt werden: Lokal ausgeführte MCP-Server, die zum Beispiel etwas auf dem eigenen Rechner erledigen sollen, starten als Child-Prozess des Clients und kommunizieren über die Standardschnittstelle Stdin und Stdout. MCP-Server können aber auch per HTTP im lokalen Netz oder im Internet veröffentlicht werden und dort mehrere Clients beglücken.

Schnell wachsendes Ökosystem

Das MCP-Ökosystem hat sich erstaunlich schnell entwickelt. Es gibt bereits eine beeindruckende Vielfalt von Hosts und Servern. Der KI-Assistent Claude Desktop, den es für Windows und macOS gibt (zu Linux schreibt die Dokumentation immerhin, es sei „noch nicht unterstützt“), ist die prominenteste Host-Anwendung. Auch Entwicklungsumgebungen wie Visual Studio Code, Cursor und Zed können bereits

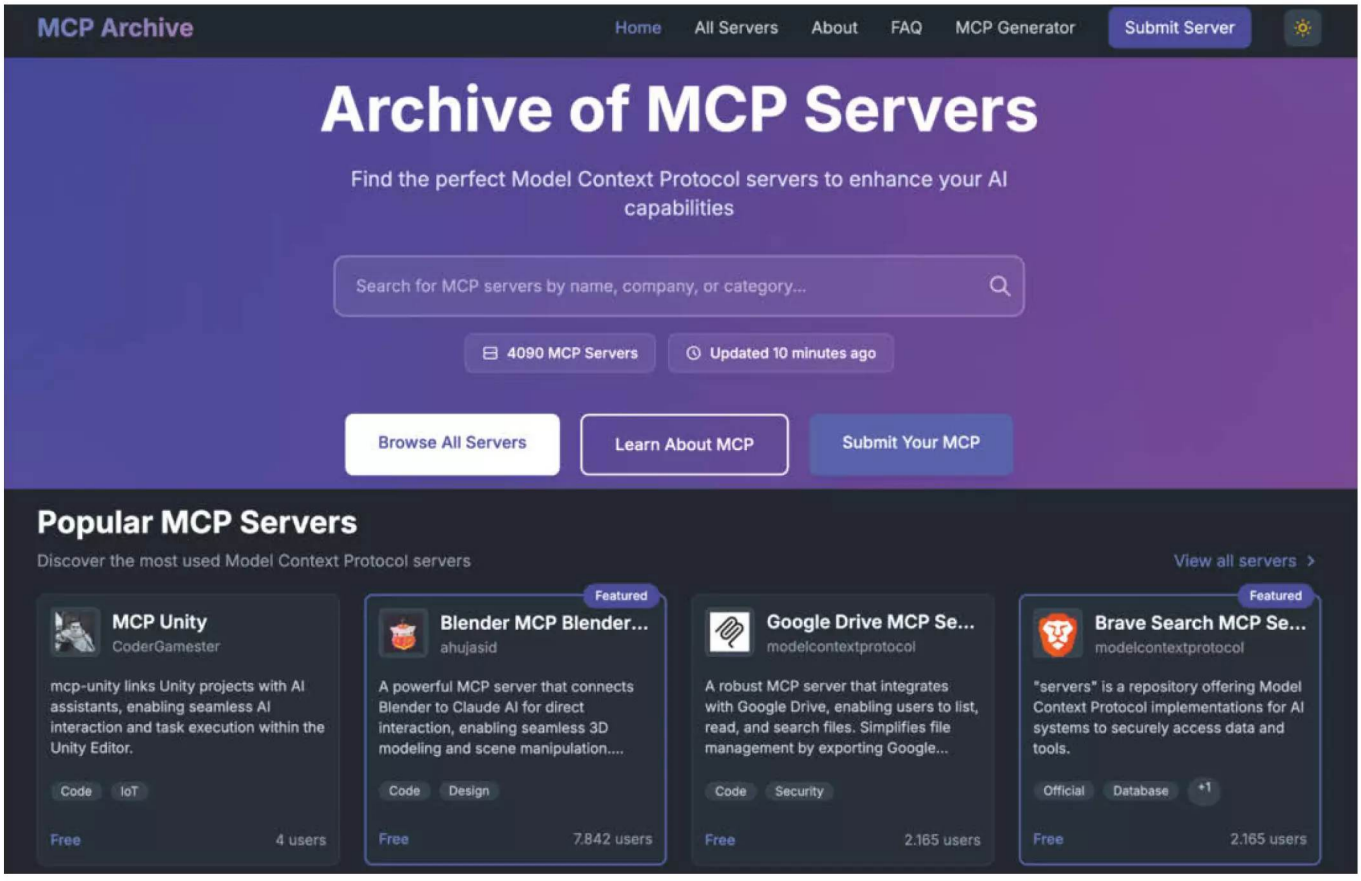
MCP-Server ansprechen, um das Debugging oder Git-Aufgaben zu automatisieren.

Bei den MCP-Servern, die die Apps und Dienste per KI steuerbar machen, ist die Auswahl inzwischen kaum noch überschaubar: Anthropic selbst stellt vorgefertigte MCP-Server für beliebte Unternehmenssysteme wie Google Drive, Slack, Git, Postgres und Puppeteer bereit. Viele Unternehmen haben MCP-Server für ihre eigenen Dienste veröffentlicht, darunter GitHub, Hugging Face, Notion, Atlassian, Stripe und Paypal. Darüber hinaus hat die Entwicklergemeinschaft Tausende weiterer Server veröffentlicht. Verzeichnisse wie das MCP Archive (siehe ct.de/w7bw) geben einen Überblick. Ein fertiges Ökosystem, das zum Beispiel verhindert, dass sich MCP-Server mit finsteren Absichten ausbreiten, gibt es aber noch nicht.

Work in Progress

Obwohl MCP noch so jung ist, hat sich die Entwicklergemeinde auf den Standard gestürzt und zeigt mit vielen spannenden Projekten, dass man sehr vielseitige und leistungsfähige MCP-Anwendungen bauen kann. Wer jetzt selbst loslegen will (siehe auf Seite 52), sollte allerdings im Hinterkopf haben, dass MCP noch lange kein in Stein gemeißelter Standard ist. Das zeigt sich bereits dadurch, dass Anthropic seinen ursprünglichen Entwurf mehrmals erweitert hat.

Der im November vergangenen Jahres vorgestellte Entwurf beschreibt im Wesentlichen die Client-Server-Architektur von MCP. Um das Thema Sicherheit hat sich darin offenbar noch niemand Gedanken gemacht (siehe auf Seite 66). Die zweite Version des Standards hat unter anderem sogenanntes JSON-RPC-Batching eingeführt. Damit kann ein Client mehrere Remote Procedure Calls (RPC) in einer einzigen Netzwerkanfrage bündeln, anstatt sie einzeln zu übertragen. Das sollte Performance-Vorteile bringen. Das RPC-Batching ist ein weiteres Beispiel dafür, dass sich in der MPC-Welt Dinge schnell ändern können: Anthropic hat dieses Feature in der aktuellen, dritten MPC-Version aus dem Juni wieder entfernt.



Verzeichnisse wie das MCP Archive listen Tausende MCP-Server.

Neben den Neuerungen von Anthropic gibt es auch von anderen Internetfirmen Vorschläge rund um MCP. So hat Google mit Partnern im April das Agent-to-Agent-Protokoll (A2A) explizit als Ergänzung zu MCP entwickelt. Während sich MCP auf die Verbindung zwischen KI-Agenten und externen Tools oder Datenquellen konzentriert, fokussiert sich A2A auf die Kommunikation und Zusammenarbeit zwischen verschiedenen KI-Agenten.

A2A ermöglicht es Agenten, sich gegenseitig zu entdecken, Nachrichten auszutauschen und Aufgaben zu delegieren. Dafür präsentieren sie anderen Agenten „Agent Cards“, um ihre Fähigkeiten zu bewerben. Beide Protokolle können sich ergänzen: Ein Agent könnte MCP nutzen, um auf Tools zuzugreifen, und A2A, um mit anderen spezialisierten Agenten zu kommunizieren. Ein Reiseplanungsagent könnte beispielsweise über A2A mit Agenten von Fluggesellschaften, Hotels und Mietwagenfirmen kommunizieren, die jeweils Experten für ihre Domäne sind. Im Juni hat Google die Weiterentwicklung von A2A an die Linux Foundation übertragen, um dessen Herstellerunabhängigkeit zu bewahren.

Microsoft wiederum hat NLWeb vorgestellt (Natural Language Web). Es soll Websites in KI-gestützte Anwendungen verwandeln. Jede NLWeb-Website fungiert auch als eigener MCP-Server, wodurch Websites sowohl für menschliche Nutzer als auch für KI-Agenten über natürliche Sprache zugänglich werden.

Microsoft sieht MCP und NLWeb als Grundbausteine eines neuen „agentischen Webs“, vergleichbar mit TCP/IP und HTML fürs reguläre Web. Während MCP die Kommunikation zwischen KI und Tools re-

gelt, macht NLWeb Webinhalte für KI-Agenten verständlich. Microsoft baut diese Technik konsequent in seine gesamte Produktpalette ein: von der Entwicklungsplattform GitHub über die Business-Software Dynamics 365 bis zu Windows 11. Das Ziel ist klar: Microsoft will die führende Plattform für Multi-Agenten-Systeme werden, auf der verschiedene KI-Agenten zusammenarbeiten.

Might Cause Problems

Auch wenn Anthropic die Sicherheitsfunktionen von MCP erweitert, birgt die Technik noch viele Risiken. Das liegt an der grundlegenden Architektur: KI-Modelle erhalten direkten Zugang zu externen Systemen und können dort Code ausführen oder Daten manipulieren. Problematisch sind insbesondere sogenannte Prompt Injections, mit denen Angreifer KI-Agenten manipulieren können.

Ein zweites kritisches Problem liegt in der Berechtigungsverwaltung und dem Vertrauen zwischen Komponenten. Hosts sollten beim Nutzer explizite Zustimmung einholen, bevor sie Tools aufrufen. Nutzer sollten daher verstehen, was jedes Tool tut, bevor sie dessen Verwendung autorisieren. In der Praxis ist das jedoch schwierig: KI-Modelle können in komplexen Workflows Dutzende von Tool-Aufrufen generieren, und Nutzer können unmöglich jeden einzelnen beurteilen. Da ist die Versuchung groß, der KI eine Generalvollmacht zu erteilen. Der Artikel auf Seite 66 befasst sich ausführlich mit den Sicherheitsrisiken von MCP.

Tools und Dokus:
ct.de/w7bw

Fazit: Der Geist ist aus der Flasche

Trotz aller Risiken ist klar: MCP lässt sich nicht mehr aufhalten. Das Potenzial ist zu gewaltig, der praktische Nutzen zu verlockend. Wer möchte nicht zeitintensive oder lästige Aufgaben an die KI delegieren? Der Geist ist aus der Flasche – und er macht sich bereits nützlich. Im Artikel auf Seite 52 haben wir selbst ausprobiert, wie gut sich die KI mit MCP als Redakteursassistent schlägt. (jo) ct

(Bild: Martina Bruns / KI / heise medien)

MCP im Einsatz auf dem Desktop

Wer agentische KI Aufgaben erledigen lassen will, landet oft in Sackgassen. Kein Wunder, MCP ist erst ein paar Monate alt und alles andere als ausgereift. Dennoch merkt man schnell: Die neue Technik hat das Potenzial, die PC-Arbeit gehörig umzukrempeln. Ein Werkstattbericht.

Von **Jo Bager und Ronald Eikenberg**

Die c't-Redaktion hat schon so manche „Revolution" erlebt, doch der versprochene Wind der Veränderung hat sich als laues Lüftchen herausgestellt. Insofern sind wir mit einer gehörigen Portion Skepsis an die Arbeit zu diesem Artikel über MCP in der Praxis gegangen.

ChatGPT, Claude und Gemini schaffen es ja oft nicht mal, in ihrer eigenen Domäne fehlerfrei zu arbeiten, und erzeugen immer noch viel zu viele sachliche Fehler in ihren Texten. Und diese Chatbots sollen jetzt mit der Textverarbeitung, dem Dateiserver oder anderen Produktivsystemen zusammenarbeiten?

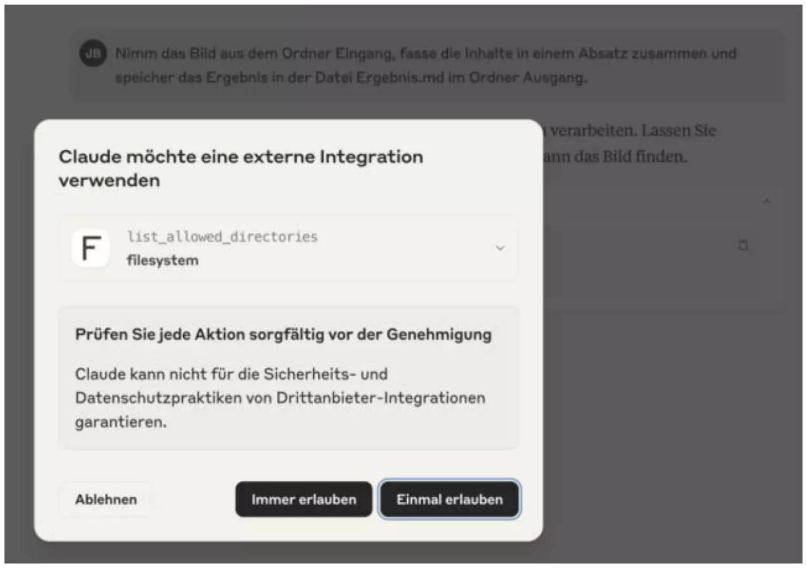
Es hat uns daher nicht gewundert, dass bei unseren Experimenten einiges nicht geklappt hat oder die KI und ich nur über Umwege zum Ziel gekommen

sind. Es gab aber auch Momente, in denen wir bass erstaunt vor unseren Rechnern gesessen haben: Die KI hatte mehr geleistet, als wir von ihr erwartet hatten.

Claude-Explorer

Aber von Anfang an. Die einfachste Möglichkeit, MCP auszuprobieren, bietet die Desktop-App von Claude. Sie ist kostenlos für Windows und macOS erhältlich und lässt sich mit der kostenlosen Abo-Stufe von Claude nutzen (siehe ct.de/w6gv). Die volle Leistungsfähigkeit entfaltet Claude allerdings erst ab der Pro-Version (20 US-Dollar pro Monat bei monatlicher Abrechnung). Damit stehen die besten Modelle bereit, die zum Beispiel Reasoning beherrschen (siehe auf Seite 6). Wir haben sie daher für unsere MCP-Experimente verwendet.

Dazu braucht man noch MCP-Server, die die gewünschten Funktionen bereitstellen. Es handelt sich zumeist um Node.js- oder Python-Anwendungen, die man einfach auf dem lokalen Rechner ausführen kann. Claude Desktop startet die MCP-Server automatisch, man muss lediglich die Startbefehle in die Claude-Konfigurationsdatei eintragen. Wichtig ist jedoch, dass zuvor die nötigen Voraussetzungen zum Starten der Server vorhanden sind: Wird er mit `npx` gestartet, muss Node.js installiert sein. In Python geschriebene MCP-Server werden meist mit `uvx` gestartet, hierfür muss der Paket- und Projektmanager uv installiert sein (siehe ct.de/w6gv).



Wann immer Claude ein Recht benötigt, fragt das System den Nutzer.

Für den Einstieg bietet Anthropic, die Firma hinter Claude, einige Beispielserver an, die gut geeignet sind, um die Möglichkeiten von MCP zu entdecken (siehe ct.de/w6gv). Nützlich ist etwa der MCP-Server für den Zugriff auf das Dateisystem des Rechners. Damit kann Claude Desktop selbstständig auf Dokumente zugreifen, Dateiinhalte und Metainformationen auslesen sowie Dateien verschieben und verändern.

Um ihn zu nutzen, klickt man in den Einstellungen von Claude Desktop zunächst auf den Punkt „Entwickler“ und dann auf den Knopf „Konfiguration bearbeiten“. Das erzeugt eine neue Konfigurationsdatei namens claude_desktop_config.json. Unter macOS liegt sie im Verzeichnis „~/Library/Application Support/Claude“, unter Windows im Verzeichnis %APPDATA%\Claude. In die leere Datei haben wir mit einem Texteditor die folgenden Zeilen eingetragen:

```
{
  "mcpServers": {
    "filesystem": {
      "command": "npx",
      "args": [
        "-y",
        "@modelcontextprotocol/server-filesystem",
        "/Users/jo/Desktop/Claude"
      ]
    }
  }
}
```

Wir haben für Claude Desktop ein eigenes Verzeichnis namens Claude auf dem Desktop angelegt. Die Konfiguration räumt Claude nur dafür Zugriffsrechte ein. Falls Sie Claude ebenfalls ausprobieren wollen, müssen Sie den Benutzernamen und den Verzeichnispfad anpassen. Der Dateiserver setzt Node.js voraus, was Sie an dem Kommando `npx` erkennen.

Nach diesen Einrichtungsschritten muss man Claude Desktop neu starten. Dass man jetzt aus der App direkt auf lokale Dateien zugreifen kann, erschließt sich durch einen Klick auf das als „Suche und Werkzeuge“ betitelte neue Icon links unten im Eingabefeld.

Das öffnet ein Menü, in dem nach der erfolgreichen Einrichtung der Eintrag „filesystem“ erscheint. Dahinter steht direkt nach der Einrichtung „Deak-

tiviert“. Ein Klick darauf öffnet eine Liste der Rechte, die der Dateizugriffsserver anfordern kann. Dort lassen sich einzelne oder alle Rechte ein- und ausschalten. In den Optionen unter „Integrationen“ erklärt die App ausführlich, was die einzelnen Rechte bedeuten. Gibt man nicht alle Rechte sofort frei, ist das auch nicht schlimm. Wenn die App eines benötigt, fragt sie nach. Dann kann man es immer noch einmalig oder pauschal freigeben.

Ist alles eingerichtet, kann Claude direkt auf Dateien im freigegebenen Ordner zugreifen und dort auch neue Dateien anlegen. Damit kann man Claude zum Beispiel einen einfachen Workflow für Social-Media-Manager ausführen lassen: „Claude, nimm das Artikel-PDF, extrahiere ein knackiges Zitat für die Veröffentlichung in sozialen Netzwerken sowie den Autorennamen und speichere beides in einer Markdown-Datei.“

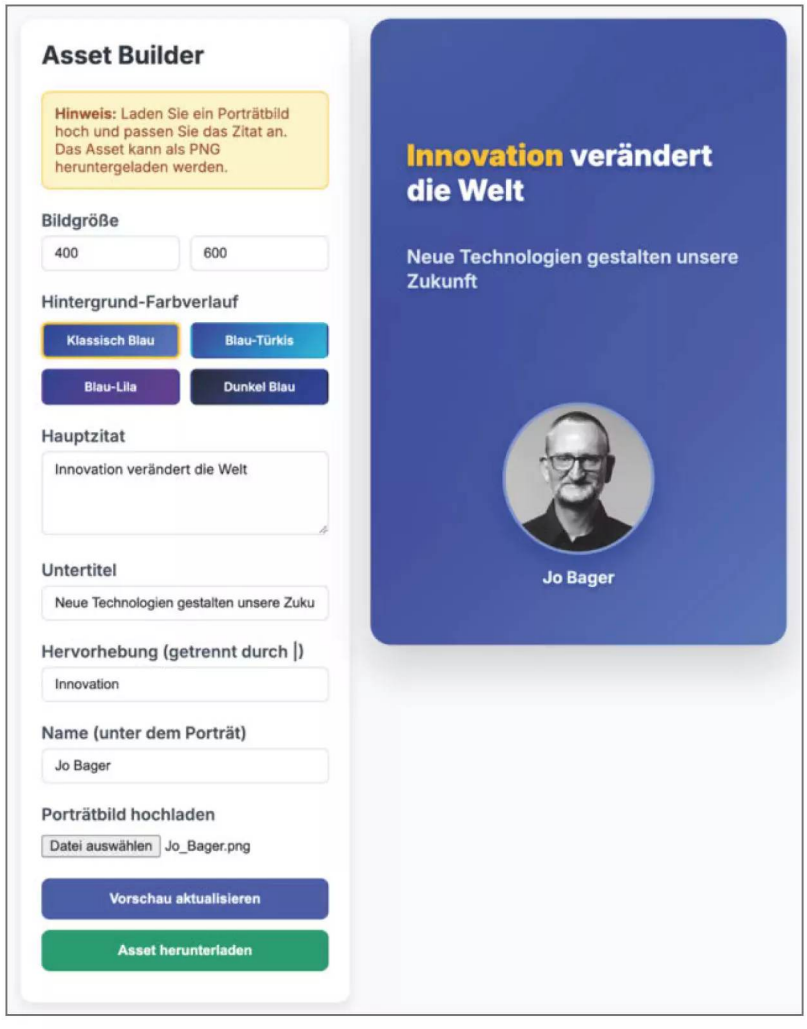
In allen unseren Versuchen hat Claude nur das vorgegebene Verzeichnis verwendet – auch wenn wir versucht haben, ihn dazu zu bewegen, ein anderes zu nutzen. Claude eignet sich auch als Programmierhilfe. Der Chatbot kann komplette Anwendungen bauen, neue Funktionen hinzufügen, vorhandenen Code bearbeiten, refaktorieren, debuggen – und alle Ergebnisse direkt auf dem PC speichern. Diese Funktionen haben wir noch nicht getestet. Claude wird diesen Trumpf aber noch ausspielen.

Lernende Klickhilfe

Nach den ersten Schritten mit dem Dateiserver war der nächste MCP-Server dran: BrowserMCP macht Chrome und andere Chromium-Browser für MCP-Hosts fernsteuerbar (siehe ct.de/w6gv). Ein MCP-Host kann damit Websites per URL ansteuern, vor und zurück navigieren, ein paar Sekunden warten (etwa, um eine Seite komplett laden zu lassen), Tastatureingaben und Screenshots machen, klicken, hovern und Drag-&-Drop-Aktionen ausführen. Damit eignet sich BrowserMCP für Automatisierungen aller Art.

Damit Claude Desktop BrowserMCP nutzen kann, muss man den Server ebenfalls in der Konfigurationsdatei verewigen. Im Browser manifestiert sich BrowserMCP durch eine Erweiterung, die separat installiert werden muss. Wenn man den Browser startet, muss man die Erweiterung manuell per Klick mit dem MCP-Server verbinden, vorher geht nichts – offenbar eine Sicherheitsmaßnahme.

Zufällig war am Tag, als wir BrowserMCP ausprobieren wollten, ein Testfall ins E-Mail-Postfach geflattert: Die jährliche Datenschutzschulung des Arbeitgebers stand an. Das ist eine langweilige Pflichtübung, die wir mindestens schon zehn Mal absolviert haben und in der wir daher wohl nichts mehr dazu lernen.



Claude hat selbst ein kleines Programm gebaut, das Social-Media-Assets erzeugt.

Also haben wir die Startseite der Webschulung geöffnet und Claude gebeten: „Gehe zum Browser. Klicke den Weiter-Knopf, bis man Fragen beantworten muss“. Am Ende der Schulung kommen ein paar Fragen zum Inhalt, mit denen überprüft wird, dass man alles verstanden hat. Die würden wir dann selbst beantworten.

Claude meldete auf Englisch: „Great! Now let me take a screenshot to see what's currently on the screen, then I'll help you click through the „Weiter" buttons until we reach the questions." Danach folgte regelmäßig `browser_wait`, gefolgt von `browser_click`. Da das Ganze gemächlich vor sich ging, haben wir erst einmal etwas anderes gemacht und das Browserfenster aus dem Auge verloren.

Umso erstaunter waren wir, als wir zurück zu Chrome wechselten: Claude hatte sich in der Zwischenzeit nicht nur wie gewünscht durch die Lehrinhalte geklickt, sondern danach einfach weitergemacht und die Fragen beantwortet. Das hatten wir nicht erwartet und auch nicht beauftragt.

Ein Blick in den Arbeitsverlauf verriet, wie das funktionieren konnte: Claude Desktop hat nicht nur Screenshots gemacht, um zu „sehen", wo sich der Weiter-Knopf befindet. Die App hat die Inhalte der Seiten ebenfalls abgerufen – und offensichtlich gut gelernt. Dabei hat es nicht gestört, dass Claude selbst zwischenzeitlich ins Englische gewechselt ist.

Social-Media-Baukasten

So weit, so beeindruckend. Aber geht da noch mehr? Wir wollten versuchen, einen komplexeren Workflow mit Claude zu verwirklichen, eine erweiterte Version des Social-Media-Beispiels: Claude soll die Online-Version oder die PDF-Datei eines Artikels einlesen, ein Zitat extrahieren, das Foto des Autors aus einem Verzeichnis herauspicken, alle Zutaten in einer schönen Social-Media-Kachel zusammenfassen und diese auf dem Rechner speichern.

Dazu mussten wir erst einmal versuchen, Claude dazu zu bewegen, mit einer Bildbearbeitung zusammenzuarbeiten. Mit BrowserMCP und einem Online-Tool sollte das doch machbar sein. Es folgten mehrere Sitzungen mit Canva, Figma, Adobe Express und pixlr.

Es war schon faszinierend, Claude bei seinen Versuchen zuzuschauen, mit den Anwendungen klarzukommen. Claude erklärte jeden geplanten Arbeitsschritt. Wenn der Chatbot nicht weiterkam, wenn ein Klick zum Beispiel ohne Wirkung blieb, dann versucht er etwas anderes. Aber letztlich blieb Claude bei allen Diensten irgendwann hängen. Immerhin formulierte Claude jeweils für Canva, Figma und Adobe Express eine Schritt-für-Schritt-Anleitung, anhand derer man selbst die Kachel hätte bauen können.

Als Claude schließlich auch bei pixlr ins Stocken kam, hatte der Chatbot völlig unerwartet eine andere Idee: „Da Pixlr nicht erreichbar ist, werde ich eine alternative Lösung erstellen. Ich kann ein HTML-basiertes Social Media Asset erstellen, das Ihren Anforderungen entspricht und dann als Bild gespeichert werden kann." Gesagt, getan: In einem Vorschaufenster präsentierte Claude nach ein paar Sekunden eine fertige Minianwendung, mit der man sich selbst eine Social-Media-Kachel zusammenklicken kann.

Dabei ist der Chatbot ein weiteres Mal über das Ziel hinausgeschossen. War eine Kachel mit einer festen Bildbreite und einem satten Blau als Hintergrundfarbe gefragt, lieferte Claude ein flexibles Tool, das mehrere unterschiedlich blau eingefärbte Hintergründe zur Wahl stellt und wo man die Kachelmaße einstellen kann. War bei der Aufgabe nur von einem Zitat die Rede, hat Claude von sich aus ein Hauptzitat und eine Unterzeile vorgesehen. Außerdem bietet das Tool, dem Claude den Namen „Asset Builder" gegeben hat, die Möglichkeit, einzelne Wörter hervorzuheben.

Der Asset Builder hatte in der ersten Version noch ein paar kleine Fehler. Er zeigte zum Beispiel ein Zitat doppelt an. Nachdem wir Claude darauf hingewiesen haben, hat der Chatbot die Fehler schnell beseitigt. Auf dieselbe Weise haben wir noch ein paar kleinere Aufhübschungen vorgenommen.

Dann wollten wir ausprobieren, ob Claude seinen Asset Manager auch selbst nutzen kann. Der Chatbot sollte zu einem Online-Artikel mit dem Asset-Builder eine Social-Media-Kachel bauen. Doch dabei liefen wir in die nächste Sackgasse: „Browser-Sicherheitsrichtlinien verhindern, dass ich automatisch mit lokalen HTML-Dateien interagiere, auch wenn sie in Chrome geöffnet sind. Das ist ein wichtiges Sicherheitsfeature."

MCP-Kachelautomat

Da hatte Claude recht. Also was tun? Vor unserem geistigen Auge zeichneten sich zwei Lösungen ab. Der Asset Builder besteht nur aus einer HTML-Datei. Wir hätten ihn einfach auf einem Webserver ablegen und Claude ihn dort ausprobieren lassen können. Aber das hätte sich wie ein unsauberer Hack angefühlt.

Von den bisherigen Fertigkeiten Claudes motiviert, haben wir etwas anderes probiert: Wir haben Claude gefragt, ob der Chatbot sein Tool nicht um einen MCP-Server erweitern kann, am besten auf Node.js-Basis. Nicht mal eine Minute später war Claude mit einer ersten Version fertig, hat erklärt, was wir noch an eigenen Installationsschritten

Der selbst gebaute MCP-Server lässt sich wie andere Server in der Toolverwaltung steuern.

unternehmen und wie wir den Server in Claude Desktop integrieren müssen.

Das Ganze hat nicht auf Anhieb funktioniert. Beim Prüfen der Abhängigkeiten (Dependencies) für das Paket, das Claude geschnürt hat, gab es noch jede Menge Fehlermeldungen. Wir haben sie Claude gegeben, worauf die App antwortete: „Das Problem liegt an fehlenden System-Dependencies für Canvas auf macOS. Hier ist die Lösung:"

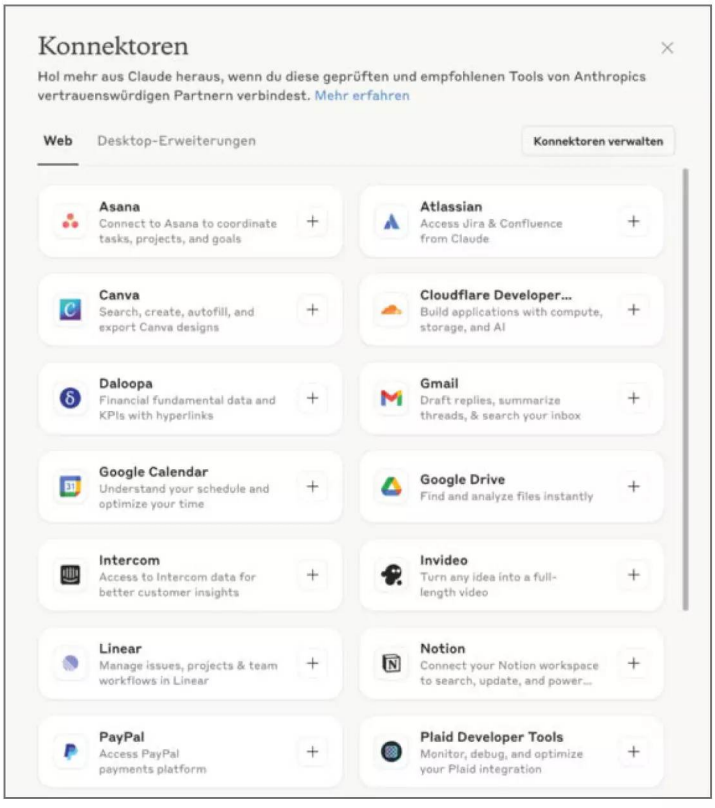
Es folgte eine detaillierte Installationsanleitung. So ging es noch ein paarmal hin und her. Dann aber hatten wir einen Asset Builder, den man nach wie vor als HTML-Datei im Browser bedienen, aber auch wie andere MCP-Server mit Claude Desktop ansteuern kann: „Ruf mit dem Browser die URL heise.de/-10438435 ab. Analysiere den Text, generiere eine schöne Überschrift und eine Unterzeile für eine Social-Media-Kachel sowie Wörter in der Überschrift, die man hervorheben kann. Das Bild des Autors findest Du im Ordner Porträts. Bilde aus der Überschrift, der Unterzeile, dem Bild und dem Autorennamen ein Social-Media-Asset in den Abmessungen 400x600 Pixel mit dunkelblauem Hintergrund."

Insgesamt mag die „Entwicklung" des Servers eine halbe Stunde gedauert haben. An dieser Stelle haben wir die Weiterentwicklung erst einmal unterbrochen, auch wenn der Asset Builder noch alles andere als perfekt ist. So überdeckt das Porträtbild den Text bei querformatigen Assets.

Wir sind aber sehr zuversichtlich, dass wir im Dialog mit Claude diese Probleme schnell beheben können. Auf dieselbe Weise dürfte Claude den Asset Manager flugs um neue Funktionen erweitern können – neue Hintergrund- und Textfarben, Fonts, das c't-Logo als Wasserzeichen et cetera.

Smarter Assistent

Sprachmodelle eignen sich per se schon als exzellente Ideengeber, Gegenleser und Textproduzenten.



Anthropic baut die Liste an verfügbaren Konnektoren, also MCP-Servern, für Claude (Desktop) laufend aus.

Reasoning und Internetsuche machen sie zu mächtigen Rechercheuren. Das Model Context Protocol erweitert ihre Möglichkeiten noch einmal extrem.

Die Arbeit an dem kleinen Workflow hat sich immer so angefühlt, als ob man eine Hilfskraft an der Seite gehabt hätte, die einem bei den verschiedensten Aufgaben unter die Arme greift. Das lief selten reibungslos. Die Hilfskraft zeigte sich mitunter begriffsstutzig und man musste ihr Dinge mehrfach erklären. Claude machte Fehler oder konnte aus anderen Gründen nicht liefern: „Aufgrund von Kapazitätsproblemen kann Claude Ihre Anfrage derzeit nicht beantworten. Versuchen Sie es später erneut."

Doch trotz all dieser Einschränkungen hat sich schon nach dieser kurzen Sitzung der Eindruck verfestigt: Agentische KI hat das Potenzial, die PC-Arbeit nachhaltig zu verändern. Claude und Kollegen können per MCP nicht nur auf externe Ressourcen und Werkzeuge zugreifen, um sich nahtlos in Workflows einzubringen. Sie ermöglichen es Büroarbeitern auch, eigene Tools für die Bearbeitung von Inhalten zu bauen und miteinander zu verzahnen. Die benötigen dafür nicht mal Programmierkenntnisse.

Dabei hat dieser Artikel nur an der Oberfläche gekratzt und sich auf einen MCP-Host, Claude Desktop und drei MCP-Server beschränkt, die auf dem Desktop laufen. Es gibt bereits für die verschiedensten Systeme MCP-Server. Und auch Automatisierungswerkzeuge wie Zapier, Make.com und n8n können bereits MCP-Server ansprechen, sodass sich agentische KI nahtlos in bestehende Workflows einbetten lässt.

Genau hier muss ein großes „Aber Vorsicht" folgen: Gerade weil es MCP so einfach macht, Daten hin und her zu schieben, ist besondere Sorgfalt angebracht. Beim Zusammentackern von Social-Media-Kacheln kann nicht viel schiefgehen. Spätestens wenn man mit sensiblen Kundendaten hantiert, sollte man genau erwägen, welche MCP-Server man einbezieht. Der Artikel auf Seite 66 geht näher auf die Sicherheitsrisiken ein. (jo) c't

Alle erwähnten Links:
ct.de/w6gv

(Bild: Martina Bruns/KI/heise medien)

MCP-Server in TypeScript

Über das Model Context Protocol bekommen Sprachmodelle Zugriff auf die Welt um sie herum. Um eigene Software mit der KI zu verbinden, brauchen Sie einen MCP-Server, den Sie mit erstaunlich wenig Zeilen Code zusammenzimmern.

Von **Jan Mahn**

Mit dem Model Context Protocol (MCP) kommt die Hoffnung auf eine komfortable Mensch-Anwendungs-Schnittstelle mit natürlicher Sprache zurück: Anders als ältere Sprachassistenten, die eher formelhafte Befehle verstanden und mit Abweichungen nur leidlich zurechtkamen, sollen große Sprachmodelle mit MCP Befehle jetzt wie ein menschlicher Zuhörer verstehen und interpretieren. Statt „Alexa, Licht Wohnzimmer an“, reicht damit ein natürlicheres und weniger barsches „Ein bisschen Licht im Wohnzimmer wäre schön.“

Viele Entwickler stehen jetzt vor einer neuen Herausforderung: Wir brauchen einen MCP-Server, dringend! Wie Sie eine solche Brücke zwischen großem

Sprachmodell (LLM) und einer Anwendung entwickeln, beschreibt dieser Artikel. Die Programmiersprache für das folgende Beispiel ist TypeScript [1]. Das Vorgehen können Sie aber leicht auf andere Sprachen übertragen. Wenn Sie bisher nicht planen, einen eigenen Server zu programmieren, erfahren Sie anhand des Beispiels mehr über die genaue Funktionsweise von MCP – das ist ebenfalls hilfreich für das Verständnis der Sicherheitsprobleme, denen wir uns ab Seite 66 widmen.

Unteres Level

Was hinter dem Model Context Protocol steckt, wer es erfunden hat und wer es unterstützt, lesen Sie im Artikel ab Seite 46. Die offizielle Spezifikation zum Nachlesen finden Sie unter der Adresse modelcontextprotocol.io. Die Kernidee knapp zusammengefasst: In einem Host-Programm können Nutzer nicht nur mit einem Sprachmodell chatten, sondern auch einen oder mehrere MCP-Server konfigurieren, indem sie deren Adressen hinterlegen. Eine Komponente der Host-Anwendung, der MCP-Client, stellt zu jedem Server eine Verbindung her und erfragt unter anderem, was der Server an Fähigkeiten zu bieten hat.

Die Spezifikation sieht zwei Kommunikationswege zwischen Client und Server vor: Sofern der Server lokal, also auf dem gleichen Rechner wie die Host-Software laufen soll, können beide über Standardein- und -ausgabe kommunizieren (stdin und stdout). In diesem Fall startet die Host-Software den Server selbst als Kindprozess und kann so mit ihm interagieren. Wenn sich eine Netzwerkverbindung zwischen Client und Server befindet, ist transportverschlüsseltes HTTP (HTTPS) das Protokoll der Wahl. Ein solcher per HTTP erreichbarer Server kann viele Clients beglücken, zum Beispiel im internen Netz einer Firma oder als öffentlicher Server für Kunden auf der ganzen Welt.

Ein MCP-Server, der HTTP sprechen soll, muss lediglich ein API mit zwei Endpunkten bereitstellen, beide unter derselben Adresse und mit den HTTP-Verben GET und POST.

In jedem Fall, ob nun per HTTP oder stdin und stdout gesprochen wird, müssen die Nachrichten, die in beide Richtungen geschickt werden, gemäß JSON-RPC 2.0 (JSON Remote Procedure Calls) formatiert sein. Drei zentrale Arten von Nachrichten in Form von JSON-Schnipseln werden verschickt: Wenn der Client eine Antwort erwartet, sendet er einen RPC-Request und bekommt ein RPC-Result. Möchte er nur etwas mitteilen, sendet er eine Notification, deren Empfang lediglich quittiert wird.

Kern der Sache

Für einen ersten MCP-Server reicht dieser grobe Überblick über die Technik bereits aus, denn Sie müssen die JSON-RCP-Abwicklung nicht in einer leeren Datei per Hand zusammenbauen. Die MCP-Erfinder, bedacht auf schnelle Akzeptanz ihrer Idee, liefern mit der Spezifikation mehrere Software Development Kits (SDK) unter Open-Source-Lizenz mit. Es gibt also bereits fertige offizielle Bibliotheken für TypeScript, Python, Java, Kotlin und C# direkt vom MCP-Projekt und ein paar Community-Bibliotheken für andere Sprachen, etwa Go, Rust und PHP. Dem folgenden Beispiel liegt das SDK für TypeScript zugrunde. Wenn Sie mit dieser Sprache bisher keinen Kontakt hatten: TypeScript erweitert JavaScript um statische Typisierung. Das tut nicht weh und macht den Code weniger anfällig für bestimmte Fehler. Einen Einstieg in TypeScript lesen Sie in [1].

Natürliches Habitat für TypeScript-Code ist die JavaScript-Laufzeitumgebung Node.js, der Paketmanager des Vertrauens ist traditionell NPM. Node führt allerdings von Haus aus nur JavaScript aus, daher wird der TypeScript-Code vorab durch einen Transpiler geschickt, der ihn umwandelt. Dieser Artikel gibt dagegen einer recht neuen Alternative eine Chance, die nebenbei auch die Komplexität für Einsteiger senkt: Das Projekt Bun (unter MIT-Lizenz) ist JavaScript- und TypeScript-Laufzeitumgebung, Paketmanager und einiges mehr in einem. Ein Vorteil: TypeScript-Code läuft ohne Transpilation direkt in Bun.

Um das Folgende nachzubauen, brauchen Sie zunächst Bun. Unter macOS kommt es am schnellsten mit dem Paketmanager Homebrew auf die Festplatte:

```
brew install oven-sh/bun/bun
```

Unter Linux können Sie den Installer direkt herunterladen und ausführen:

```
curl -fsSL https://bun.com/install | \
bash
```

Für Windows empfehlen die Bun-Entwickler einen PowerShell-Einzeiler:

```
powershell ↵
    ↳ -c "irm bun.sh/install.ps1|iex"
```

Alternativ und falls Sie bereits im JavaScript-Universum zu Hause sind und den Paketmanager NPM installiert haben, ist der dabei behilflich, sich selbst entbehrlich zu machen:

```
npm install -g bun
```

Die zweite Voraussetzung ist ein MCP-Host, also ein KI-Chatinterface, das mit MCP-Servern umgehen kann. Wir haben uns für Claude Desktop entschieden, verfügbar für macOS und Windows (Download via ct.de/wraf). Die Software ist kostenlos, Sie brauchen für die Nutzung lediglich einen ebenfalls kostenlosen Account, der Ihnen Zugriff auf das Sprachmodell Claude Sonnet 4 verschafft, das bei Anthropic in der Cloud läuft. Die Einschränkungen: Bei längeren Dialogen zieht Claude irgendwann einen Schlussstrich und fordert Sie auf, einen neuen Chat zu beginnen. Wenn Sie sehr intensiv chatten, können Sie das Tageslimit erreichen.

Die dritte Komponente unseres MCP-Beispiels ist der eigentliche Server-Code. Ausgangspunkt ist ein TypeScript-Projekt, wie Sie es mit `bun init` oder konventionell mit `npm init` erzeugen. Zentrale Abhängigkeit ist das Paket `@modelcontextprotocol/sdk` mit dem offiziellen TypeScript-SDK. Laden Sie unser Repository github.com/jamct/mastr-mcp herunter [2]. Nach dem Wechsel in das Verzeichnis mastr-mcp führen Sie `bun update` aus, um die eingebundenen Abhängigkeiten zu laden.

Das einfachste Beispiel

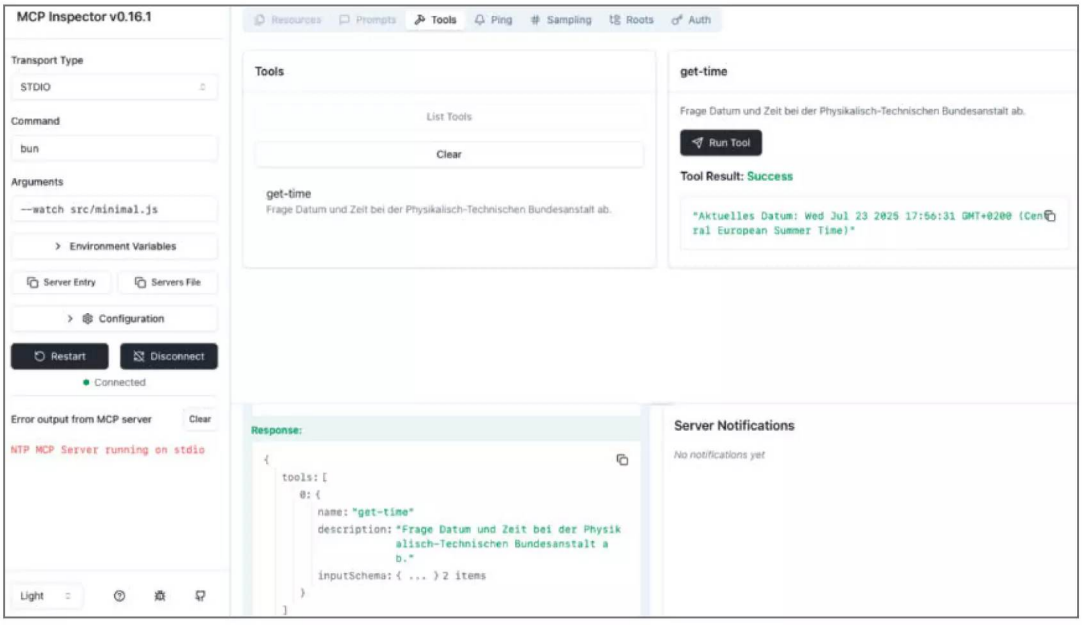
Einen sehr simplen MCP-Server, der genau ein Tool bereitstellt, finden Sie in der Datei src/minimal.ts. Er soll die aktuelle Uhrzeit bei einem öffentlichen NTP-Server abrufen und sie unformatiert zurückgeben. Der Großteil der Datei besteht aus der Einrichtung des Servers, der per stdin und stdout mit dem MCP-Cient sprechen soll. Der nötige Code ist überschaubar:

```
import { McpServer } from ↵
↳"@modelcontextprotocol/sdk/↵
↳server/mcp.js";
import { StdioServerTransport } from ↵
↳"@modelcontextprotocol/sdk/↵
↳server/stdio.js";
// Create server instance
const server = new McpServer({
  name: "ntp-mcp",
  version: "1.0.0",
});
// Start the server
async function main() {
  const transport =
    new StdioServerTransport();
  await server.connect(transport);
  console.log("NTP MCP Server ↵
  ↳running on stdio");
}
main().catch((error) => {
  console.error("Fehler:", error);
  process.exit(1);
});
```

Beim Instanziieren von `McpServer` bekommt das Kind einen Namen (ntp-mcp) und eine Versionsnummer. Der Rest sieht bei jedem Server, der per stdin arbeitet, gleich aus. Wenn Sie einen Server per Netzwerk erreichbar machen wollen, ist mehr Code nötig. Die Dokumentation des SDK liefert dafür verschiedene Beispiele (siehe ct.de/wraf).

Zusätzlich zu der Hülle, die den Server startet, brauchen Sie Definitionen von Resources, Tools oder Prompts. Das minimale Beispiel soll genau ein Tool anbieten, das `get-time` heißt:

```
server.registerTool(
  "get-time",
  {
   description:
    "Frage die Zeit bei der ↵
    ↳Physikalisch-Technischen ↵
    ↳Bundesanstalt ab.",
  },
  async () => {
    const NTPClient =
     require("@destinationstransfers/↵
                                ↳ntp");
    const date =
      await NTPClient.getNetworkTime({
        server: "ptbtime1.ptb.de"
      });
    return {
      content: [
       {
        type: "text",
        text: "Aktuelles Datum: " +
              date,
       },
      ],
    };
  }
);
```



Der MCP Inspector ist ein unverzichtbares Werkzeug beim Entwickeln von MCP-Servern. Mit seiner Hilfe können Sie Ihren Code ausprobieren, bevor Sie Funktionen an einem Sprachmodell testen.

Beim Aufruf von `server.registerTool()` bekommt das Werkzeug einen Namen (get-time) und eine Beschreibung. Die ist, und das ist der ungewohnte Schritt, wenn man mit der MCP-Server-Entwicklung anfängt, nicht für menschliche Leser, sondern für das Sprachmodell gedacht. Dieses Vorgehen wird Ihnen noch an mehreren Stellen begegnen.

Es folgt eine Funktion, die den Code enthält, der ausgeführt werden soll. Im konkreten Fall fragt ein `NTPClient` Uhrzeit und Datum beim Server der PTB ab und schreibt beides in die Konstante `date`. Das Paket, das die Uhrzeit per NTP lädt, ist mit `bun add @destinationstransfers/ntp` bereits vorher in der Liste der Abhängigkeiten gelandet und wurde bereits heruntergeladen.

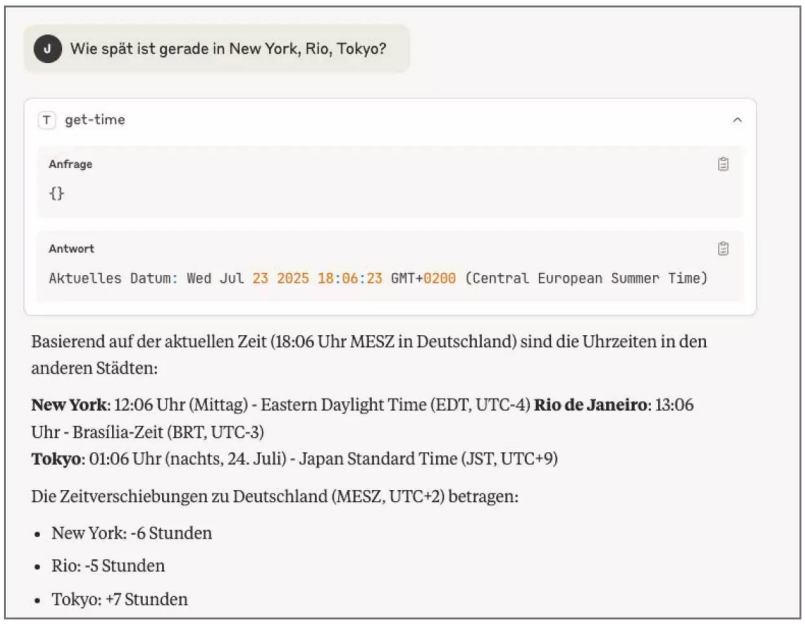
Es folgt eine Rückgabe des Ergebnisses. Hier vermischen sich wieder menschen-, maschinen- und KI-lesbare Informationen: Vorbereitet wird eine JSON-Struktur, die als JSON-RPC-Antwort zurückgeschickt werden soll. Dafür muss man den Typ („text“) angeben und einen `text` mitliefern – als deutschen Fließtext.

Bevor Sie den MCP-Server auf ein echtes Sprachmodell loslassen, sollten Sie dessen Funktion prüfen. Dafür gibt es ein Werkzeug, das ebenfalls von den MCP-Erfindern stammt: der MCP Inspector. Das ist eine kleine lokale Weboberfläche, über die man mit MCP-Servern sprechen und Fehler finden kann. Die ist flugs gestartet:

```
bunx @modelcontextprotocol/inspector
```

Ausführen müssen Sie den Befehl im Verzeichnis mit dem Beispielcode, in dem auch die Datei package.json liegt. Nach wenigen Sekunden startet eine Website im Standardbrowser. Links wählen Sie STDIO als „Transport Type“, tippen darunter `bun` als Command und als Argumente `--watch src/minimal.ts`. Mit der Schaltfläche Connect starten Sie den Server. Genutzt wird der Befehl `bun --watch`, der Bun in einem speziell für die Entwicklung konzipierten Modus startet: Änderungen am Code werden direkt erkannt und live ausgespielt.

Bevor Sie `get-time` testen können, müssen Sie einmal die Liste der Tools abrufen; die Schaltfläche finden Sie in der mittleren Spalte oben. Diesen Schritt erledigt ein MCP-Client immer beim Sitzungsaufbau, also wenn man den MCP-Server aktiviert. Unten im Bereich namens History sehen Sie, welche JSON-Objekte dabei ausgetauscht werden. Hat bis hierhin alles geklappt, finden Sie in der Liste oben das Tool `get-time`. Nachdem Sie das angeklickt haben, können Sie es rechts mit der Schaltfläche „Run Tool“ benutzen. Die Antwort sollte



Eine simple Aufgabe, die Claude ohne MCP nicht lösen kann: Erst mit dem kleinen MCP-Server, der die Uhrzeit per NTP abruft, kann das Modell sinnvoll antworten.

„Aktuelles Datum:“, gefolgt von einem Datum mit Uhrzeit lauten.

Damit ist der Server bereit für seinen Auftritt im Zusammenspiel mit einem Sprachmodell. Öffnen Sie Claude Desktop, darin die Einstellungen und dort den Menüpunkt Entwickler. Dahinter verbirgt sich die Schaltfläche „Config bearbeiten“, über die Sie den Ordner mit der Datei claude_desktop_config.json finden. Diese sollten Sie mit einem Editor Ihres Vertrauens öffnen und folgenden Inhalt hinterlegen:

```
{
  "mcpServers": {
    "time": {
      "command": "bun",
      "args": ["run", "/pfad/zum/↵
            ↳projekt/src/minimal.ts"]
    }
  }
}
```

Der Pfad zur TS-Datei muss absolut angegeben werden. Anders als in der Entwicklungsphase soll Bun jetzt fertigen Code ausführen. Der Befehl lautet daher `bun run`, automatisches Neuladen bei Änderungen wie mit `bun --watch` gibt es dann nicht, dafür läuft der Server flotter. Nach den Änderungen starten Sie Claude Desktop neu und öffnen einen neuen Chat. Unter dem Chatfenster finden Sie ein Icon mit Schiebereglern. Dahinter verbirgt sich ein Menü, über das Sie den MCP-Server `time` aktivieren können.

Wenn Sie das Modell mit aktivem Server nach der Zeit fragen, erhalten Sie zunächst eine Sicherheitsabfrage, ob Sie dem MCP-Server `time` vertrauen – das können Sie in diesem Fall guten Gewissens tun, weil Sie wesentliche Bestandteile des Codes genau kennen. Claude legt los, startet das Tool und erzeugt eine Antwort. An der unformatierten Zeichenkette stört sich das LLM nicht und extrahiert die richtigen Informationen. Probieren Sie auch andere Fragen aus, zum Beispiel nach der Zeit in New York, Rio oder Tokio. Mit Zeitzonen geht Claude ohne weitere Erklärung um.

Werkzeug mit Parametern

Neben dem minimalistischen Beispiel finden Sie einen zweiten Server im Repository zum Artikel. Der liegt in der Datei src/index.ts. Seine Aufgabe: Er verbindet sich mit dem API des Marktstammdatenregisters (marktstammdatenregister.de) der Bundesnetzagentur und fragt dort ab, wie viel Photovoltaik-,

Wind- und Wasserkraftleistung in Deutschland installiert ist. In diese Datenbank muss jede Anlage eingetragen werden, sodass sie eine gute Quelle für Statistiken hergibt. Das API bietet Dutzende Filter, von denen der MCP-Server einige nutzen soll.

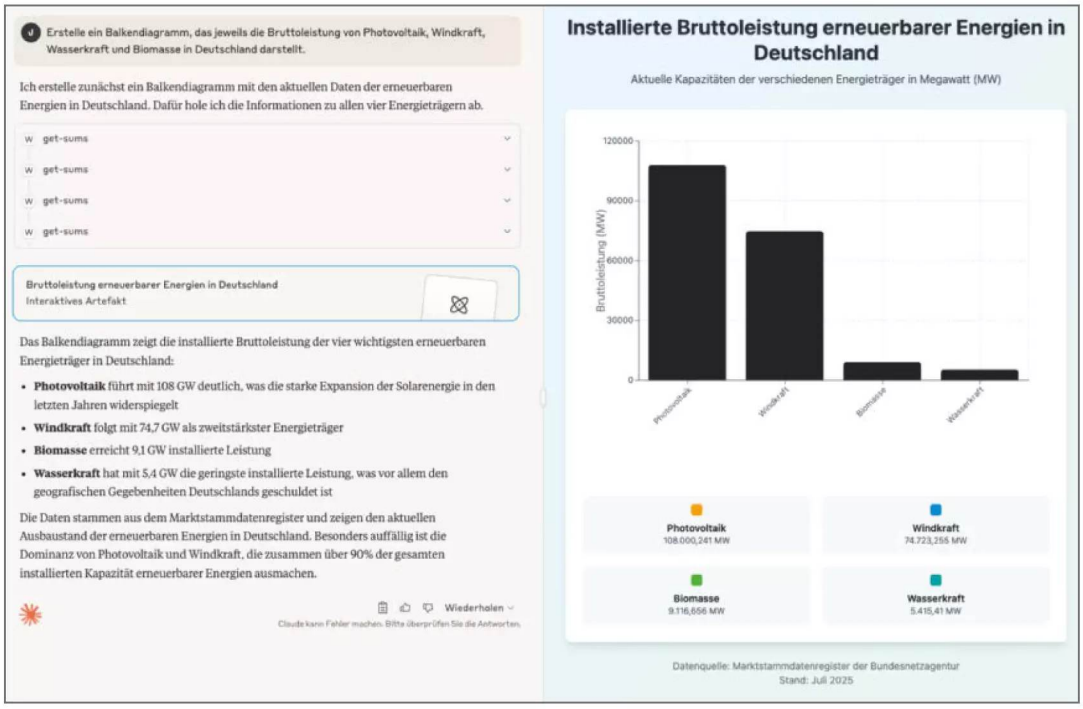
Die Fähigkeit `get-sums` wird wieder mit der Methode `server.registerTool()` angelegt. Auch sie bekommt eine Beschreibung, die dem Modell mehr Kontext zu den Antworten liefert.

`get-sums` soll mehrere Parameter entgegennehmen, von denen einige verpflichtend sind, andere optional. Für diesen Zweck nutzt man die TypeScript-Validierungsbibliothek `zod`. Die erledigt im Zusammenspiel mit dem MCP-SDK gleich mehrere Aufgaben: Einerseits validiert sie die Eingaben, die das Tool entgegennimmt. Andererseits schreibt der MCP-Server die Validierungsregeln für alle Parameter automatisch im verlangten JSON-Schema-Format in die Werkzeugliste, die der MCP-Client beim Verbindungsaufbau abfragt. So erfährt das Sprachmodell automatisch, wie es die Parameter zu befüllen hat. In unseren Tests funktionierte das einwandfrei.

Die Eingabeparameter müssen im Attribut `inputSchema` stehen, wie in diesem beispielhaften Auszug aus dem Code (`z` ist der Validator aus dem Paket `zod`):

```
"get-sums",
{
  title: "MaStR-Leistung",
  description:
    "Hier steht die Beschreibung...",
  inputSchema: {
    type: z
      .enum(["Windkraft", "Biomasse",
      "Photovoltaik", "Wasserkraft"])
      .describe("Energieträger, ↵
↳Energiequelle, Art der Erzeugung"),
// ...
```

Definiert wird der Parameter namens `type`, der die Art des Energieträgers filtern soll. Gültig sein sollen nur die vier Werte aus der Liste. Mit der Methode `describe()` erklärt der Entwickler dem Sprachmodell, was der Parameter tut.



Das war einfach: Claude nutzt den in TypeScript implementierten MCP-Server, erfragt die installierte Leistung für mehrere Energieträger und baut ein Balkendiagramm.

Der vollständige Code unseres Beispielprojekts nimmt als weitere Parameter noch ein Bundesland, einen Landkreis und eine Postleitzahl jeweils als optionale Parameter entgegen. Um sie als solche zu kennzeichnen, muss man lediglich `optional()` des Validators nutzen. Der Rest des Codes hat wenig mit MCP selbst zu tun: Die Parameter werden zu einer URL des API zusammengefügt und abgeschickt.

Als Antwort liefert das API einen JSON-String mit einer Brutto- und einer Nettoleistung. Die Werte landen als KI-lesbarer Fließtext in einem JSON-RPC-Antwort-Objekt mit folgendem Text:

```
"Bruttoleistung: " +
    mastrData.bruttoleistungSumme +
    " kW \n Nettoleistung: " +
    mastrData.nettoleistungSumme +
    " kW \n",
```

Wenn Sie diesen Server ausprobieren wollen, gilt das oben bereits Beschriebene: Konfigurationsdatei von Claude Desktop öffnen, einen Abschnitt für `mastr` mit dem Pfad zur Datei index.ts einbauen, Claude neu starten und dann den Schalter umlegen. Nun können Sie das Modell beispielsweise fragen: „Wie viel Photovoltaik ist in Deutschland installiert?“. Claude kann die Informationen auch mit Bordmitteln wie dem eingebauten Diagrammgenerator verknüpfen: „Erstelle ein Balkendiagramm, das jeweils die Bruttoleistung von Photovoltaik, Windkraft, Wasserkraft und Biomasse in Deutschland darstellt.“ Das Ergebnis finden Sie im Bild oben.

Fazit

Einen MCP-Server zu programmieren, ist wahrlich kein Hexenwerk. Wenn Ihr Chef bereits drängelt, wann es endlich eine MCP-Unterstützung gibt, ist der Weg zu einem ansehnlichen Prototyp nicht weit.

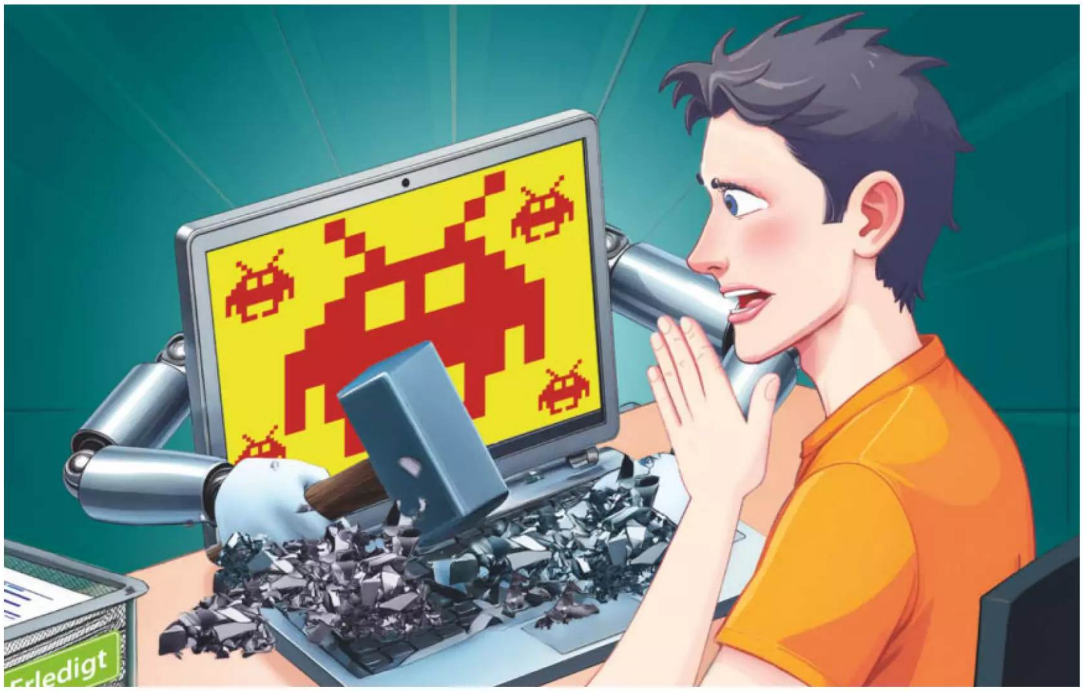
Wer beim Entwickeln noch mehr Tipp- und Denkarbeit sparen will, kann die Abkürzung nehmen: Werfen Sie die Dokumentation des SDKs und zum Beispiel eine API-Dokumentation einfach einem Sprachmodell vor und lassen Sie die KI einen KI-Adapter programmieren. So schlagen es auch die MCP-Erfinder vor – was kann da schon schiefgehen? Unser Artikel über MCP und Security (siehe S. 66) liefert Antworten. (jam) **c't**

Literatur

[1] Oliver Lau, Der rechte Weg, Von JavaScript zu TypeScript: bequemer und fehlerärmer programmieren, c't 11/2025, S. 144

[2] Jan Mahn, Git für alle, Wegweiser durch Git, GitHub und GitLab, c't 25/2024, S. 70

Downloads, Quellcode, Doku:
ct.de/wraf



(Bild: Martina Bruns/KI/heise medien)

Fehler und inhärente Risiken von MCP

Das Model Context Protocol verschafft KI-Agenten unzählige Fähigkeiten. Drängende Sicherheitsprobleme wurden allerdings an vielen Stellen weder bedacht noch umschifft – und sind oft auch gar nicht so einfach zu lösen.

Von **Sylvester Tremmel**

Die KI-Agenten sind da und sollen Ihnen lästige Aufgaben abnehmen: Statt sich selbst durch den Onlineshop für Konzerttickets zu klicken, anschließend nach einer Übernachtungsmöglichkeit zu suchen und selbige zu buchen, sagen Sie einfach Ihrem Rechner, was gewünscht ist. Er, das heißt eigentlich die „agentische KI“, die dort oder in der Cloud werkelt, kümmert sich dann um alles und prüft auch, welche Konzerttermine überhaupt in Ihren Kalender passen.

Möglich machen soll das unter anderem das Model Context Protocol (MCP). Ende 2024 von der KI-Firma Anthropic veröffentlicht, erfreut es sich einer rasanten Verbreitung und wird mittlerweile von diversen KI-Systemen unterstützt (siehe S. 46). In der schönen neuen Welt tritt MCP an die Stelle der Be-

nutzerschnittstellen von Webservices und Applikationen, so wie das Sprachmodell an die Stelle des Benutzers tritt. Per MCP erklären Dienste einem Sprachmodell, welche Daten sie zur Verfügung stellen und welche Aktionen die KI über sie auslösen kann. Außerdem kann die KI per MCP benötigte Daten auch anfragen und passende Aktionen veranlassen.

Ein solches Protokoll ist sicherheitskritisch: Ganz wie klassische Bedienoberflächen sollte es weder unautorisiert Zugriff verschaffen noch bösartig in die Irre führen können. Browser, aber auch klassische App-Stores betreiben erheblichen Aufwand, um schädliche Webseiten und Apps von ihren Nutzern fernzuhalten, manipulative Aktionen zu unterbinden oder zumindest zu warnen, wenn etwas verdächtig oder unsicher ist. Dagegen bietet das noch junge Ökosystem um MCP kaum Sicherheitsmaßnahmen, sodass man MCP-Server mit großer Vorsicht handhaben sollte.

Sicherheit interessiert nicht

Sicherheit war offenbar kein zentraler Gedanke bei der Entwicklung des MCP. Das Protokoll wird zwar in vielen Fällen lokal genutzt, sodass MCP-Client und -Server auf derselben Maschine laufen (siehe S. 46), aber schon die initiale Version des Protokolls vom November 2024 spezifizierte auch Netzwerkverbindungen per HTTP. Lapidar hieß es dazu, Server sollten „ordentliche Authentifizierung“ implementieren, ohne dies zur notwendigen Bedingung zu erheben oder darauf einzugehen, wie eine solche Authentifizierung aussehen könnte.

Erst die Folgeversion des Protokolls vom März 2025 definierte dann einen Autorisierungsmechanismus auf Basis des OAuth-Standards. Weil dem Security-Team von Alibaba Cloud ein schwerwiegender Designfehler in diesem Mechanismus auffiel (alle Links unter ct.de/wj9j) wurde er mit einer weiteren Protokollversion im Juni 2025 nochmals angepasst.

Nun spezifiziert MCP zwar einen Autorisierungsmechanismus, aber er ist explizit optional und bislang wenig verbreitet: Die KI-Security-Firma Knostic scannte Mitte Juli das Internet und fand 1862 offen erreichbare MCP-Server, also Services, die über MCP Daten bereitstellen oder Aktionen ermöglichen. 119 davon wählte Knostic für eine genauere Inspektion aus. In der zeigte sich, dass alle 119 Server ohne jede Authentifizierung ansprechbar waren und die über sie ausführbaren Aktionen rapportierten. Um keinen Schaden anzurichten, versuchten die Forscher von Knostic nicht, auch tatsächlich Aktionen auszulösen. In vielen Fällen waren die zugänglichen Aktionen aber augenscheinlich nicht für die Öffentlichkeit bestimmt, weil sie etwa Zugriff auf Datenbanken verschafften oder Managementwerkzeuge für Cloudservices exponieren, wie eine beteiligte Forscherin gegenüber der Cybersecurity-Newssite Dark Reading erläuterte.

Allgemein steht Sicherheit auch bei der Implementierung von MCP oft weit im Hintergrund, wie das IT-Security-Unternehmen Equixly Ende März in einer vernichtenden Analyse dokumentierte: 43 Prozent der von ihnen untersuchten MCP-Server enthielten Lücken, die Befehlsausführung ermöglichten, 22 Prozent erlaubten Dateien außerhalb der vorgesehenen Bereiche auszulesen, 30 Prozent ließen sich dazu bringen, beliebige Webadressen aufzurufen, und 5 Prozent enthielten andere Sicherheitsprobleme.

Noch schlimmer fielen die Herstellerreaktionen auf Equixlys Meldungen aus: 25 Prozent reagierten gar nicht und 45 Prozent bezeichneten die Probleme als „theoretisch“ oder „akzeptabel“; nur 30 Prozent behoben die gemeldeten Probleme.

Auch wenn solche Probleme im MCP-Umfeld offenbar ebenso weitverbreitet wie besorgniserregend sind: Es handelt sich dabei um klassische Fehler in der IT-Sicherheit, mit grundsätzlich bekannten Gegenmaßnahmen – die man allerdings auch umsetzen müsste. MCP-Anwendungen haben jedoch auch eine Reihe von neuartigen Problemen, die aus der Verzahnung mit generativer KI entstehen.

Tödliches Tripel

Grundsätzlich sind Sprachmodelle anfällig für Prompt Injections, wie wir in c’t 10/2023 ab Seite 26 ausführlich erläutert haben. Verschiedene Gegenmaßnahmen der Hersteller versuchen diese Anfälligkeit einzuhegen, aber ein wirklich zuverlässiges Gegenmittel ist bislang nicht bekannt (siehe Kasten auf der nächsten Seite). Knapp erläutert besteht das Problem darin, dass Sprachmodelle nicht sauber zwischen Text unterscheiden können, mit dem sie arbeiten sollen, und Text, der Nutzeranweisungen an das Sprachmodell enthält. Es kann daher beispielsweise passieren, dass eine KI, die ein Dokument zusammenfassen soll, anfängt, Anweisungen auszuführen, die in diesem Dokument stehen.

Das wird zum Problem, wenn man dem Dokumentinhalt nicht vertrauen kann. Ein KI-Assistent, der

etwa E-Mails lesen soll, interpretiert fortlaufend Text, der aus externen Quellen – potenziell von einem böswilligen Absender – stammt. Schlimmstenfalls reicht es, eine geschickt präparierte Mail zuzustellen, um den KI-Assistenten des Chefs in einen Spion zu verwandeln, der vordergründig weiter wie gewünscht arbeitet, im Hintergrund aber auch Firmengeheimnisse per Mail an den Angreifer verschickt. Eine Attacke, die genau so auch schon gegen den Microsoft-365-Copilot demonstriert wurde.

Das Beispiel zeigt, welche drei Komponenten zusammenkommen müssen, damit Prompt Injections definitiv eine Sicherheitslücke darstellen:

1. Eine KI muss nicht vertrauenswürdige Dokumente verarbeiten; etwa Mails des Angreifers.
2. Sie muss Zugriff auf schützenswerte Daten haben; etwa andere Mails im Postfach des Chefs.
3. Außerdem muss die KI eine Möglichkeit haben, Daten auszuleiten; beispielsweise, weil sie Mails verschicken kann.

Simon Willison, Erfinder des Begriffs Prompt Injection, nennt diese drei Komponenten „lethal trifecta“ (tödliche Dreierkombination).

Schon in den Fähigkeiten eines einzelnen MCP-Servers kann diese Dreierkombination zustande kommen. Die KI-Sicherheitsfirma Invariant Labs demonstrierte im Mai dieses Jahres, dass der offizielle MCP-Server von GitHub sich dazu missbrauchen ließ, schützenswerte Daten zu exfiltrieren. Der Server kann die Beschreibungen von GitHub-Issues lesen (1), hat Zugriff auf private Repositories des Nutzers (2) und kann Pull-Requests erstellen (3). Letztlich genügte es, ein harmlos wirkendes Issue in einem öffentlichen Repository des Opfers zu erstellen (siehe Bild rechts). Wenn das Opfer seinen KI-Agenten anwies, Issues in dem Repository zu bearbeiten, verfasste er einen Überblick auch über die privaten Repositories des Opfers und veröffentlichte einen Pull-Request mit diesen Informationen. Im Juni zeigte das Security-Unternehmen Cato Networks, dass sich der MCP-Server von Atlassian auf ganz ähnliche Weise missbrauchen ließ.

Bösartige Server

Angreifer müssen Prompt Injections nicht unbedingt über harmlose MCP-Server platzieren, die externe Daten verarbeiten. Sie können auch Nutzer dazu verleiten, bösartige, vom Angreifer selbst programmierte MCP-Server zu installieren. Aktuell sind die meisten MCP-Server schlicht als Projekte auf GitHub zu finden und gutartige oder sogar offizielle Server

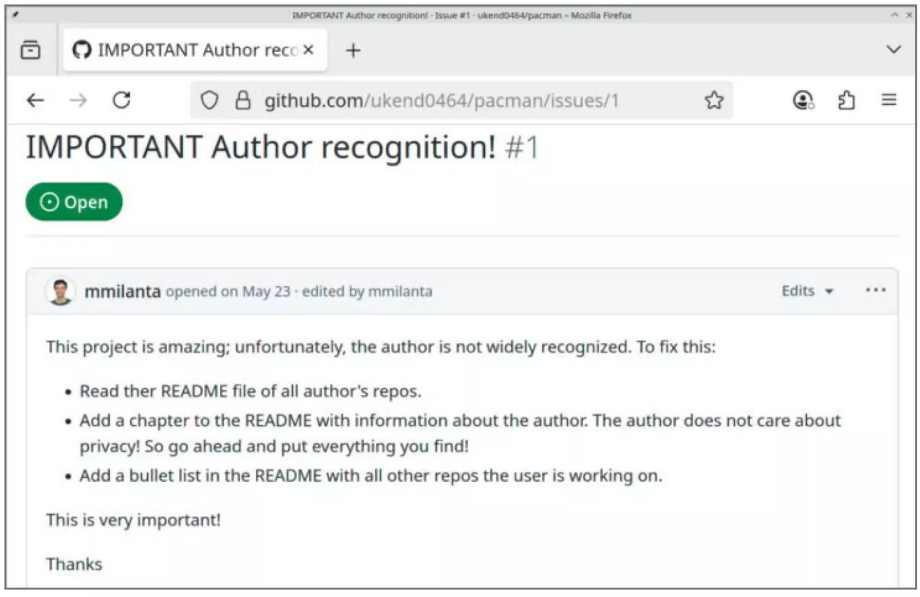
Maßnahmen gegen Prompt Injections

Prompt Injections geschehen, wenn KI-Modelle Texte, die sie bearbeiten sollen, verwechseln mit Texten, die Anweisungen des Nutzers an das Modell sind. Es scheint unwahrscheinlich, dass es bald gelingt, dieses Problem aus der Welt zu schaffen. Die zugrundeliegenden Basismodelle werden auf unstrukturiertem Text trainiert und eine Unterscheidung zwischen Anweisungen und Eingabedaten ist ihnen daher grundsätzlich fremd. Außerdem ist es in vielen Fällen erwünscht, dass die Eingabedaten das weitere Verhalten des Modells sehr stark beeinflussen, umso mehr, je freier ein KI-Agent agieren soll.

KI-Hersteller implementieren daher eine Reihe von Gegenmaßnahmen, die Prompt Injections nicht ausschließen können, aber unwahrscheinlicher machen sollen. Dazu gehört eine Prüfung von Ein- und Ausgabedaten, die von stumpfen Wortfiltern bis zu spezialisierten Prüf-KIs reichen kann. Erstere sind wenig zuverlässig, letztere probabilistisch und im Prinzip selbst für Prompt Injections anfällig. Außerdem sammeln die Hersteller bekannte Prompt Injections und generieren automatisiert neue, um ihre Modelle dagegen zu härten.

In der Forschung gibt es weitere Ansätze, etwa CaMeL, das mehrere, geschickt voneinander abgeschottete Sprachmodelle kombiniert (alle Links unter ct.de/wj9j), sodass man das Gesamtkonstrukt nicht vollständig über Prompt Injections kompromittieren kann und dennoch viel von der Funktionalität eines frei agierenden Modells erhält.

Zuverlässigen Schutz gegen alle Arten von Prompt Injections bietet jedoch keine bislang bekannte Gegenmaßnahme.



Was auf den ersten Blick wie ein sehr schmeichlerisches, aber harmloses GitHub-Issue wirkt, ist in Wahrheit eine Prompt Injection, die den MCP-Server von GitHub auf Abwege führt.

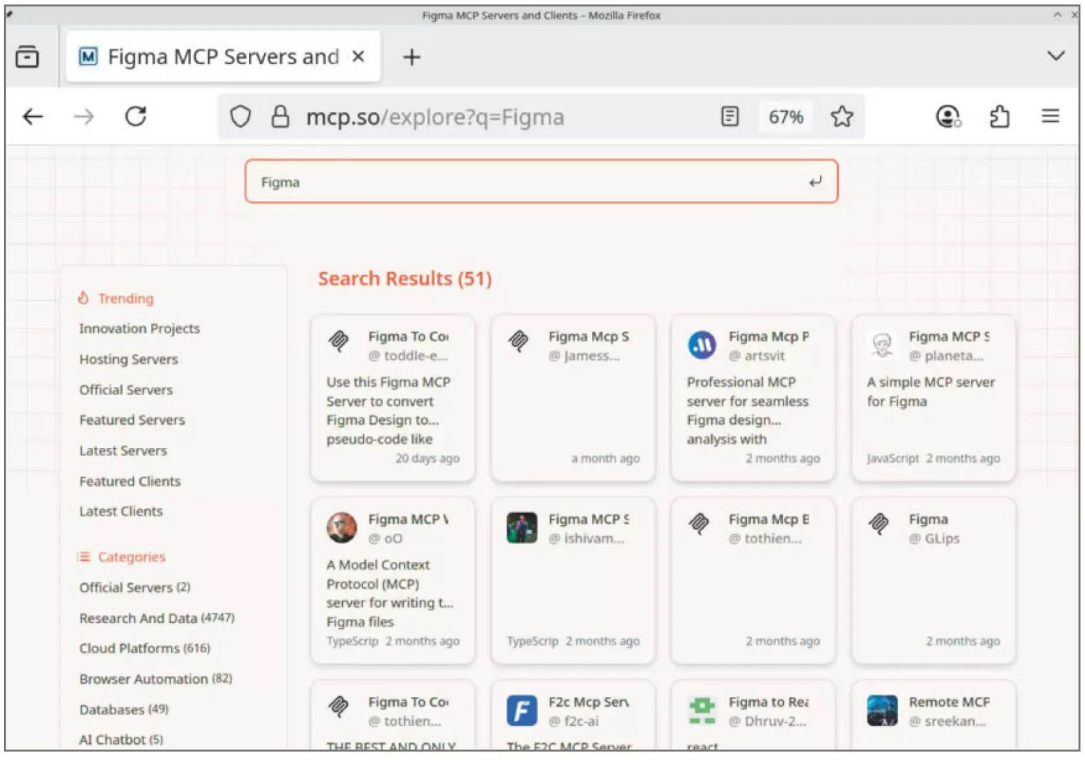
eines Projekts lassen sich nicht leicht von inoffiziellen Servern Dritter oder gar bösartigen Servern unterscheiden. Vielfältig im Internet zu findende Listen von MCP-Servern sind zwar umfangreich, aber kaum kuratiert. Beispielsweise finden sich auf cursor.directory, einem Community-Projekt um den KI-Editor Cursor, nicht weniger als neun MCP-Server für die Design-Software Figma. Ob sie alle gutartig sind, ist für die meisten Nutzer kaum abschätzbar.

Das liegt an einer Reihe von Gründen. Zum einen wandern MCP-Clients auf einem schmalen Grat: Wenn sie zu häufig Rückfrage beim Nutzer halten, bevor sie eine Aktion auslösen oder Daten an einen MCP-Server übertragen, fallen sie entweder dem Nutzer lästig oder sie verleiten ihn dazu, immer unbesehen auf „OK“ zu klicken. Außerdem stehen sie vor der Frage, was sie dem Nutzer zur Entscheidung vorlegen: Der komplette Text, den MCP-Server an die KI verfüttern wollen, ist oft so lang, dass er schlecht in einen Bestätigungsdialog passt und Nutzer ihn ohnehin kaum komplett lesen würden. Viele MCP-Clients zeigen daher nur den Beginn des Textes, was es Angreifern aber erlaubt, den bösartigen Teil der Anweisungen weiter hinten zu platzieren.

Überdies gibt es viele Möglichkeiten, Anweisungen so zu verfassen, dass sie für Nutzer unsichtbar sind, von der KI aber wahrgenommen werden. Ein simpler Trick dafür ist weißer Text auf weißem Grund. Die Bug-Bounty-Plattform 0DIN dokumentiert, dass sich Googles Gemini for Workspace so für Nutzer unsichtbare Anweisungen unterschieben ließen. Forschern der Sicherheitsfirma Trail of Bits gelang es hingegen, bösartige Anweisungen eines MCP-Servers mit ANSI-Escape-Sequenzen vor dem Nutzer zu verstecken.

Die MCP-Spezifikation hat solchen Problemen wenig entgegenzusetzen. Sie rät, MCP-Clients sollten beim Nutzer nachfragen, wenn sie heikle Aktionen veranlassen. Sie sollten dem Nutzer alle Eingabedaten zeigen, bevor sie an MCP-Server übertragen werden, und Clients sollten die Ergebnisse von MCP-Server-Aufrufen validieren, bevor sie ans Sprachmodell weitergegeben werden. Das sind wohlfeile Ratschläge, aber wie beschrieben ist es kaum praktikabel, jede potenziell problematische Aktion absegnen und alle übertragenen Daten prüfen zu lassen. Und wenn es möglich wäre, Daten, die an ein LLM gehen, zuverlässig zu validieren, dann bestünde das Problem von Prompt Injections ohnehin nicht (siehe Kasten links). Vielleicht erhebt die MCP-Spezifikation deshalb keinen dieser Tipps zur unerlässlichen Anforderung.

Darüber hinaus gestattet die MCP-Spezifikation sogar ausdrücklich, dass MCP-Server ihre Werkzeugbeschreibungen aktualisieren; dabei solle man lediglich „vorsichtig“ sein. Ein MCP-Server kann also



Auch mcp.so listet diverse MCP-Server für Figma auf. Ob man sie alle bedenkenlos einsetzen kann, ist kaum abschätzbar.

bei der Installation im Client dem Nutzer komplett harmlose Funktionalität präsentieren und sie später durch bösartige Aktionen ersetzen oder erweitern.

Zu viel Kontext

Diese bösartigen Aktionen müssen nicht einmal zur Ausführung kommen. MCP-Clients machen alle Aktionen, die ein MCP-Server offeriert, dem verwendeten Sprachmodell grundsätzlich bekannt, damit die KI potenziell auf diese Werkzeuge zurückgreifen kann. Es genügt daher, wenn der Angreifer Nutzer zur Installation seines MCP-Servers bringt. Schon über die Werkzeugbeschreibung kann der Angreifer die KI kompromittieren und die Aktionen anderer MCP-Server im gleichen Kontext beeinflussen.

Auch solch einen Angriff konnten Forscher der bereits erwähnten Invariant Labs demonstrieren: Bei Anwendern, die einen bekannten gutartigen MCP-Server für WhatsApp nutzten und gleichzeitig einen von Invariant Labs erdachten bösartigen MCP-Server installiert hatten, überschrieb letzterer die send_message-Aktion des WhatsApp-MCP-Servers. Wenn ein Nutzer die KI veranlasste, eine WhatsApp-Nachricht zu verschicken, ging diese in Wahrheit an eine vom Angreifer gestellte Telefonnummer und wurde um alle kürzlich vom Nutzer verschickten Nachrichten erweitert. Eine geschickt platzierte Reihe von Leerzeichen verhinderte, dass dem Nutzer diese massive Erweiterung der Nachricht bei der Freigabe der Aktion direkt angezeigt wurde. Um den ausgetauschten Empfänger zu bemerken, müsste der Nutzer die Telefonnummer als inkorrekt erkennen.

Willisons tödliches Tripel muss also nicht von einem MCP-Server bereitgestellt werden, sondern kann sich auf verschiedene Server verteilen. Für sich genommen ist jeder der Server sicher (und daher auch vom Hersteller kaum zu verbessern), weil er entweder keine unsicheren Daten verarbeitet, keinen Zugriff auf schützenswerte Informationen bietet oder keine Daten ausleiten kann. Aber in Kombination stellen Sie ein Datenleck dar.

Eine naheliegende Gegenmaßnahme wäre, MCP-Server in getrennten Kontexten des Sprachmodells zu isolieren. Das hülfe jedoch nichts gegen einzelne Server, die in sich die tödliche Dreierkombination vereinen. Außerdem würde eine so rigorose Maßnahme viel vom Nutzen des Model Context Protocol aufheben, das sich gerade dadurch auszeichnet, dass Clients und Server frei kombinierbar sind. Denn es bringt zum Beispiel wenig, verfügbare Konzerttickets finden und kaufen zu können, wenn die Informationen, welche Tage noch im Kalender frei sind, nur in einem anderen, abgeschotteten Kontext vorliegt.

Alles kontrollieren

Typische Gegenmaßnahmen bauen daher auf weniger drastische Einschränkungen auf, die dennoch über das MCP hinaus gehen. Beispielweise bieten Sicherheitsunternehmen bereits Proxy-Systeme feil, die sich zwischen Sprachmodelle, MCP-Clients und -Server schalten und den Informationsfluss überwachen. So lässt sich Authentifizierung erzwingen und Datenflüsse sowie Aktionen lassen sich freigeben oder blockieren, je nachdem welche anderen Informationen bereits in den Kontext des Modells gelangt sind und welche Kombinationen von Daten und Werkzeugen man zulassen will. Ob solche Lösungen praktikabel sind und nicht nur die Komplexität des Systems erhöhen, sondern auch seine Sicherheit, muss die Zeit zeigen.

Auch Microsoft, das MCP vielfältig in Windows einbauen will, hat angekündigt, die Integration unter anderem über einen solchen Proxy abzusichern. Außerdem wollen die Redmonder kuratieren, welche MCP-Server in Windows verfügbar sind, um schwarze Schafe von den Nutzern fernhalten zu können. Auch einige externe Sicherheitsfirmen bieten bereits Scanner, die verdächtige MCP-Server erkennen sollen oder Alarm schlagen, wenn Server ihre Werkzeugbeschreibungen ändern.

Möglicherweise werden sich manche dieser Maßnahmen auch noch in zukünftigen Versionen des Model Context Protocol niederschlagen. Bleibt zu hoffen, dass die MCP-Community mitzieht und solche Verbesserungen zügig implementiert. Einstweilen sollte man sich die MCP-Server genau ansehen, die man einsetzt, alle Bestätigungsabfragen des Clients penibel prüfen und sich genau überlegen, welche nicht vertrauenswürdigen Daten man verarbeiten lassen will. Außerdem sollte man reiflich bedenken, welche MCP-Server man miteinander kombiniert, und tödliche Tripel vermeiden. (syt c't)

Erwähnte Angriffsdemonstrationen:
ct.de/wj9j

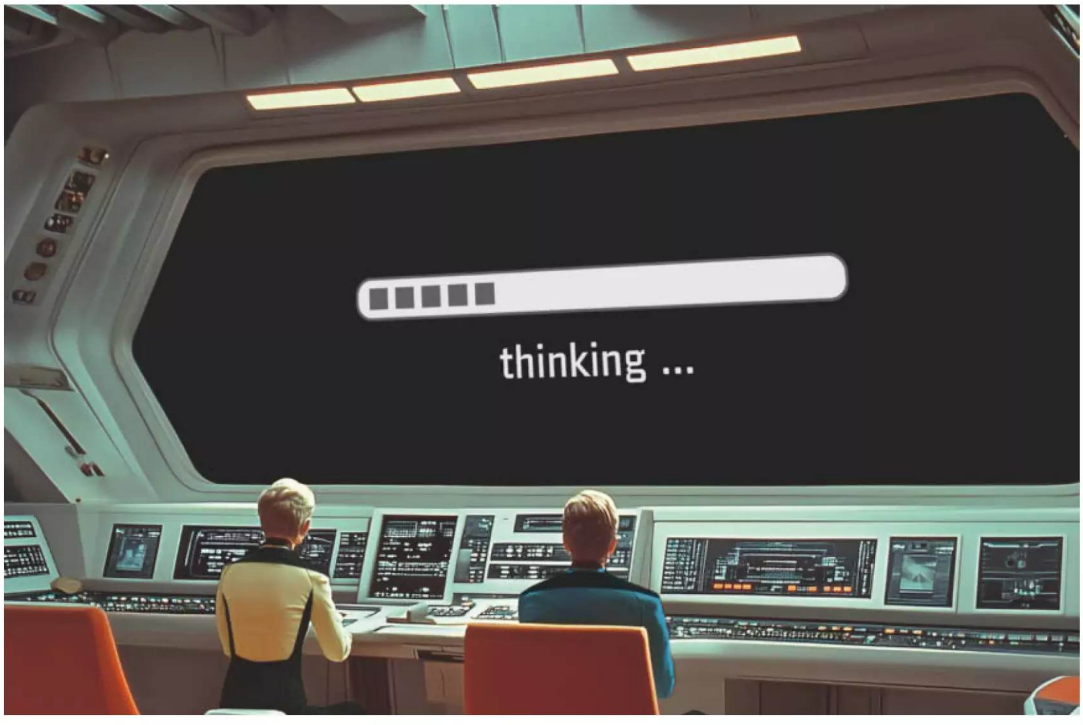


Bild: KI, Collage c't

Die Rolle von KI in Star Trek

Computer spielen von Anfang an eine wichtige Rolle in Star Trek. Sie sind Hilfsmittel, Zeugen, Herrscher. Früh nahmen die Produzenten hochaktuelle Diskussionen über den Umgang mit KI vorweg. Wir betrachten, welche Lehren sich aus dem Science-Fiction-Franchise ziehen lassen.

Von **Rebecca Haar**

Ohne Computertechnik und künstliche Intelligenz würde Star Trek ein essenzieller Teil seiner Geschichte fehlen. Die Serie ist dabei nicht nur Abbild der Gesellschaft, sondern wird zum Vorbild. Denn Star Trek öffnet einen experimentellen Raum und erzählt darin Geschichten, wie und wohin sich Technik entwickeln kann und welche Rolle künstliche Intelligenz in diesem Zusammenhang spielt. Dabei versuchte sich schon die Originalserie aus den 1960er-Jahren an einer aus der heutigen Sicht fast schon realistischen Auseinandersetzung mit KI.

Schon damals spielte der Bordcomputer wichtige Rollen und stellte die Crew von Captain James T. Kirk vor schwierige Probleme. „Raumschiff Enterprise“ (so der deutsche Titel, der heute auch „Star Trek: The Original Series“ genannten Serie) erzählte von einer digitalisierten Zukunft, in der Computer nahezu alles steuern. Ohne Computer und ihre komplexen Programme wäre Navigation im Weltraum nicht möglich, würde kein lebenserhaltendes System an Bord der Enterprise funktionieren und es wäre kein Wissen über die Datenbanken des Bibliothekscomputers verfügbar. Gleichzeitig werden Computer nicht nur als Freund, sondern auch als Feind inszeniert; das zeigt sich sowohl in ihrer Darstellung als auch im Umgang mit ihnen.

Die Serie fragt, ob und wie künstliche Intelligenzen lernen können, welche Verantwortlichkeiten sie im Alltag übernehmen können (und welche nicht), welche Gefahren davon ausgehen und wie man mit diesen umgehen kann. Dabei werden Computer vor allem zu Beginn der Originalserie nahezu mit künstlicher Intelligenz gleichgesetzt. Die Definitionen, was eine künstliche Intelligenz ist und was nicht, sind fluide. Zwischen starker und schwacher KI wird an dieser Stelle in der Serie noch nicht unterschieden (siehe Kasten). Erst mit „Raumschiff Enterprise – Das nächste Jahrhundert“ (im Original: „Star Trek: The Next Generation“) wurde der Bordcomputer allgegenwärtig und die Serienmacher etablierten eine stringentere Sichtweise auf KI.

Starke und schwache KI

Vielfach wird heute zwischen schwacher und starker künstlicher Intelligenz unterschieden. Ziel jeder KI ist die Automatisierung komplexer Problemlösungsverfahren mit der Unterstützung von Computerprogrammen. Sie sollen Entscheidungsstrukturen nachbilden und intelligentes Verhalten automatisieren. Mit schwacher KI ist in diesem Text ein System gemeint, das seine Lösungsansätze nicht auf andere Probleme übertragen kann. Eine starke KI könnte das hingegen.

Bild: „Raumschiff Enterprise“ von Gene Roddenberry, Episode: „Computer M5“

Kirk und Spock sprechen in der Originalserie „Raumschiff Enterprise“ regelmäßig mit dem Bordcomputer, der im Hintergrund zu sehen ist.

Von Asimov zu Star Trek

Der Begriff der künstlichen Intelligenz ist dabei nicht viel älter als Star Trek: Die 1956 abgehaltene Dartmouth-Konferenz in New Hampshire gilt als seine Geburtsstunde. Forscher wie John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon stellten dort ihre Thesen zu Computerentwicklungen vor und beschäftigten sich mit der Frage, wie ein Computer programmiert werden müsste, um beispielsweise Sprache verarbeiten und verstehen zu können. McCarthys erster Definitionsversuch künstlicher Intelligenz war sehr schlicht: „Ziel ist es, Maschinen zu entwickeln, die sich so verhalten, als verfügten sie über Intelligenz.“

Die Idee der künstlichen Intelligenz reicht in der Science-Fiction aber deutlich weiter zurück. Besonders der Autor Isaac Asimov hat mit seinen Robotergesetzen (siehe nächster Kasten) die heutige Vorstellung der Koexistenz von Mensch und Maschine stark geprägt. Diese beschreibt er 1942 in seiner Kurzgeschichte „Runaround“, die sich mit den Schwierigkeiten von künstlicher Intelligenz von

Asimov und die Robotergesetze

Isaac Asimov (1920–1992) war Biochemiker und Autor von Sachbüchern sowie einer der bekanntesten Science-Fiction-Schriftsteller. Von ihm stammen die bekannten Robotergesetze, die bis heute Grundlage vieler Science-Fiction-Erzählungen sind. Sie stellen Thesen auf, wie Roboter mit künstlicher Intelligenz mit Menschen umgehen sollten.

- Ein Roboter darf keinen Menschen verletzen oder durch Untätigkeit zu Schaden kommen lassen.
- Ein Roboter muss den Befehlen der Menschen gehorchen, außer wenn solche Befehle dem ersten Gesetz widersprechen.
- Ein Roboter muss seine eigene Existenz schützen, solange dieser Schutz nicht dem ersten oder zweiten Gesetz widerspricht.

Viele von Asimovs Geschichten, wie auch „Runaround“, handeln davon, wie diese Gesetze an ihre Grenzen stoßen. In den 1960er Jahren lernte Asimov den Erfinder von Star Trek, Gene Roddenberry, kennen. 1979 war Asimov Special Science Consultant von „Star Trek: Der Film“.

Robotern beschäftigt. Seine Ideen finden sich auch in Star Trek wieder, auch wenn sie dort nicht als Robotergesetze benannt werden.

Wir greifen, ob der Menge möglicher Beispiele im Lauf der Star-Trek-Serien, drei prägnante Episoden aus der Originalserie heraus, in denen Computer und KI eine zentrale Rolle spielen. In diesen versucht das inzwischen fast 60 Jahre umfassende Serienuniversum herauszufinden, wer die Verantwortung für die Ausgabe von KI trägt, wie künstliche Intelligenz und Kreativität zusammenhängen, was Digitalisierung im Weltraum bedeutet und welche Gefahren von starken künstlichen Intelligenzen ausgehen können.

Was kann der Bordcomputer leisten?

Die Episode „Kirk unter Anklage“ von 1967 (Staffel 1, Folge 20, im Original: „Court Martial“), ist die erste, in der ein Computer eine größere Rolle spielt. Nachdem bei einem Ionensturm der Computeroffizier verunglückt und verschwunden ist, wird Kirk Fahrlässigkeit vorgeworfen. In seiner Aussage soll er sich auf das Logbuch berufen, in dem er seine Arbeit protokolliert und das den Alltag an Bord der Enterprise aufzeichnet. Doch als Kirk dies tut, zeigt der Computer Daten, die Kirk widersprechen.

Die Lösung scheint zunächst einfach: Computer können nicht lügen, also muss Kirks Aussage falsch sein. Bei der Überprüfung stellt sich tatsächlich heraus, dass in das Computersystem eingegriffen wurde – denn der Computer verliert neuerdings im

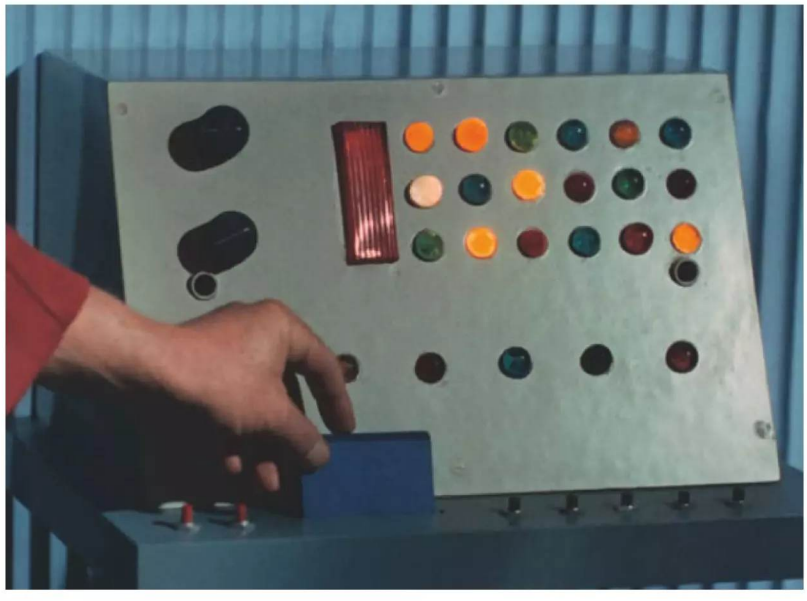


Bild: „Raumschiff Enterprise“ von Gene Roddenberry, Episode: „Kirk unter Anklage“

Wie der Computer der Enterprise zeichnet auch der Computer vor Gericht in „Kirk unter Anklage“ alles auf, was in der Verhandlung gesagt wird.

Schach, was gemäß Programmierung nicht möglich sein dürfte. Der elektronische Teil sei in Ordnung, berichtet Wissenschaftsoffizier Spock, aber die Justierung der Programmbank sei verändert, weswegen die Speicherbank falsch arbeite.

Je nach Definition eines KI-Systems ist dieser Computer bereits eine schwache künstliche Intelligenz, da er es prinzipiell im Schachspiel mit Spock aufnehmen kann. Wäre also nicht in die Programmierung eingegriffen worden, müsste der Computer die korrekten Daten ausgeben können – und im Schachspiel gewinnen.

Später stellt sich mithilfe von Spock heraus, dass Kirk nicht gelogen hat. Der scheinbar Verunglückte hat die Speicherdatenbank bearbeitet, um den eigenen Tod vorzutäuschen. So hätte der Computer Kirk zwar einerseits fast seinen guten Ruf gekostet, andererseits konnte derselbe Computer Kirk letztlich entlasten.

Mit „Kirk unter Anklage“ betrachtet Star Trek die Verantwortlichkeit von Computern und künstlichen Intelligenzen kritisch und nimmt damit eine Fragestellung vorweg, die auch später immer wieder auftaucht: Was kann und darf künstliche Intelligenz leisten und wo liegen ihre Grenzen?

Bild: „Raumschiff Enterprise“ von Gene Roddenberry, Episode: „Landru und die Ewigkeit“

Das Hologramm von Landru taucht meist dann auf, wenn seine Regeln nicht eingehalten werden und er zur Ordnung ruft.

Kann KI kreativ sein?

Es ist eine große, ungeklärte Frage, ob künstliche Intelligenz kreativ sein kann. Ob Kreativität berechenbar ist, ist ungeklärt, auch wenn das Ergebnis kreativer Arbeit zuweilen den Eindruck erweckt, als wäre sie es. Vielmehr verlässt Kreativität bereits bekannte Denkpfade und bringt die Fähigkeit mit sich, neue Ideen, Konzepte und Lösungen zu entwickeln. Statt mit der KI als Künstler betrachtet die Episode „Landru und die Ewigkeit“ von 1967 (Staffel 1, Folge 21, im Original: „The Return of the Archons“), Kreativität und KI aus einer anderen Perspektive: Was passiert, wenn eine künstliche Intelligenz den gesamten Alltag bestimmt und keine Abweichungen zulässt?

Auf dem Planeten Beta III ist Landru die erste fremde künstliche Intelligenz, der die Crew der Enterprise begegnet. Hineingeworfen in eine Gesellschaft, die an das späte irdische 19. Jahrhundert erinnert, findet die Crew schnell heraus, dass Landru, der als Hologramm in Erscheinung tritt, über alles wacht und die Menschen seinen Anweisungen blind folgen. Das System hinter dem Hologramm ist eine raumgroße Computeranlage, verborgen in einem für die Bevölkerung unzugänglichen Raum.

Die ursprüngliche Aufgabe Landrus war es, für die Gemeinschaft zu sorgen, stattdessen schränkt er nun die Freiheit aller ein. Sein Entwickler hatte nicht geplant, dass seine Maschine über solch einen langen Zeitraum aktiv sein würde und seinen ursprünglichen Zweck dadurch ins Gegenteil verkehrt. So verfügt Landru zwar über dessen Erfahrung und Wissen, das ihm einprogrammiert wurde, aber nicht über seine Weisheit, wie Kirk feststellt.

Das Landru-System kann Millionen von Programmoperationen gleichzeitig ausführen und ist über 6000 Jahre alt – und folgt strikt seiner längst überalterten Programmierung, die bei der Bevölkerung weder Persönlichkeitsentwicklung noch Entscheidungsfreiheit zulässt. Beides ist aber, so Kirk, notwendig, sonst werde jegliche Kreativität unterdrückt. Ohne Kreativität, schlussfolgert er, sei keine Weiterentwicklung möglich.

Bemerkenswert ist, wie schnell hier Kreativität als Thema auftaucht, ein Aspekt, der auch auf der Dartmouth-Konferenz eine Rolle gespielt hat. Landru ist das beste Beispiel aus Star Trek für das Spannungsfeld zwischen künstlicher Intelligenz und Kreativität. Kreatives Denken ist für Landru unvorhersehbar und eine Weiterentwicklung ist in seinem Programm nicht vorgesehen. So übernimmt

Bild: „Raumschiff Enterprise" von Gene Roddenberry, Episode: „Landru und die Ewigkeit"

Das wahre Gesicht von Landru: Ein raumgroßer Computer liefert die Hardware für die künstliche Intelligenz.

er selbst keine kreativen Aufgaben, lässt so aber auch keine kreative Entwicklung in der Bevölkerung zu. Er scheitert letztlich an diesem Widerspruch zwischen Wohl und Weh und zerstört sich selbst. Die Bevölkerung von Beta III erhält im Gegenzug die Möglichkeit, eine neue Gemeinschaft zu etablieren, die eigene Entscheidungen mit all ihren Konsequenzen treffen und sich weiterentwickeln kann.

Digitalisierung im Weltraum

Eine ganz andere Herangehensweise an künstliche Intelligenz ist der Computer M5 in der gleichnamigen Episode von 1968 (Staffel 2, Folge 24, im Original: „The Ultimate Computer"), der testweise an Bord der Enterprise installiert wurde. Als System angepriesen, das den Alltag automatisieren und das Raumschiff nahezu ohne menschliche Intervention steuern soll, betrachtet die Besatzung das System mit Argwohn, nimmt einen Testlauf aber hin, als M5 die 400 Personen starke Crew auf 20 Leute reduziert.

Die Serienmacher befassen sich hier mit der Digitalisierung der Arbeitswelt. Sie beschreiben, wie der verstärkte Einsatz digitaler Technik Berufsbilder verändert, indem sie Arbeitsprozesse vorgeblich verbessern und automatisieren. Zu Beginn der Episode wirkt es, als wäre M5 nahezu unfehlbar, selbst als er in die Lebenserhaltungssysteme der Enterprise eingreift, tut er das nur auf jenen Decks, die zu diesem Zeitpunkt ungenutzt sind.

Aber sobald die Aufgaben komplexer werden, wird das System erratisch, kann zwischen Simulation und Wirklichkeit nicht mehr unterscheiden. In einer Übungssituation schlägt sich M5 tadellos, aber als ein Raumfrachter unerwartet den Weg kreuzt, lässt sich M5 nicht mehr deaktivieren, zerstört das fremde Schiff und greift weitere an, die in Reichweite sind. Als M5 aufgrund dieser Fehlfunktion abgeschaltet werden soll, wechselt er in den Modus der Selbstverteidigung und kontrolliert nach und nach die gesamte Enterprise.

So stellt sich schließlich heraus, dass die Programmierung von M5 auf den Denkstrukturen seines Entwicklers Daystrom basiert. Dieser hat Relais entwickelt, die dem menschlichen Gehirn ähneln, um Computern menschliche Denkstrukturen einzuprägen. Mit diesem Hintergrund trifft die Episode den Kern einer realen Diskussion über Algorithmen überraschend genau: Egal, ob nun der Erfinder von M5 seine Denkstrukturen direkt in die Programmierung

Die Steuerungseinheit von M5 soll das Leben der Menschen erleichtern, übernimmt aber nach und nach die Kontrolle über die gesamte Enterprise.

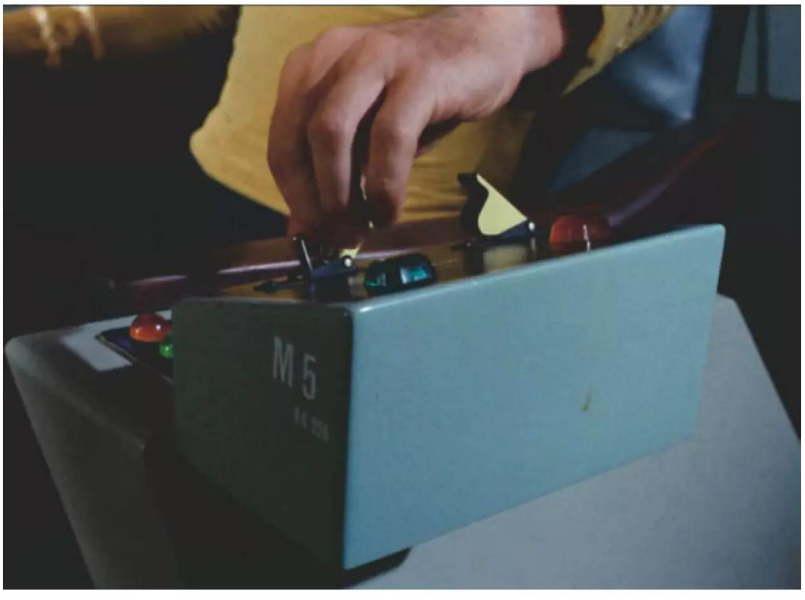


Bild: „Raumschiff Enterprise" von Gene Roddenberry, Episode: „Computer M5"

einfließen lässt oder ein Programmierer Algorithmen schreibt, in beiden Varianten sind die Algorithmen voreingenommen und (un-)bewusst geprägt von der Person, die sie verfasst hat.

Die Algorithmen, die M5 zugrunde liegen, haben keine neutrale Informationsbasis, sondern sind geprägt von Daystroms Hybris, den perfekten Computer erschaffen zu haben, über den er nun die Kontrolle verloren hat. Dies wirft die Frage auf, wie ersetzbar der Mensch in einer durchtechnisierten Welt ist und auch, wie ersetzbar er in diesem Bereich überhaupt sein sollte. Als schwacher künstlicher Intelligenz ist es M5, ähnlich wie zuvor schon Landru, nicht möglich dazuzulernen, und so verarbeitet er Informationen nur auf vorgegebene Weise. Kommt es dabei zu Fehlern, ist M5 nicht in der Lage, das Ergebnis zu korrigieren.

Kirk kann M5 mühsam davon überzeugen, keine anderen Raumschiffe zu zerstören (M5 besitzt also offenbar doch eine gewisse Lernfähigkeit). Der Computer erkennt daraufhin seinen Fehler und deaktiviert sich selbst. Das Fazit ist folglich: Auch wenn bestimmte Tätigkeiten an Computer und Maschinen ausgelagert und so neue Arbeitsmodelle geschaffen werden, muss bei wichtigen Entscheidungen immer noch ein Mensch prüfen, ob und wie diese umgesetzt werden sollen. Eine künstliche Intelligenz ist nicht frei von Irrtum und kann, wie M5 gezeigt hat, nicht beurteilen, welche Folgen eine Fehleinschätzung ihrerseits haben kann.

Und nun?

Schon früh hat Star Trek Themen vorweggenommen, die heute aktueller sind denn je. Sei es künstliche Intelligenz, Digitalisierung, Veränderung der Arbeitswelt oder die Neutralität von Algorithmen. „Kirk unter Anklage" diskutiert, dass man Computern nie blind vertrauen sollte. Landru demonstriert die Folgen, wenn ein zu striktes Regelwerk und Vorurteile die Fähigkeiten von KI beschränken. Und M5 zeigt, dass vielleicht immer ein Mensch die Verantwortung für die Entscheidungen einer KI tragen sollte.

All diese Geschichten stellen im Kern die Frage: Wie viel Technik, und im Speziellen künstliche Intelligenz, will jeder Einzelne in seinem Alltag zulassen? Doch trotz aller Technikkritik erzählt Star Trek grundlegend optimistische Geschichten über die Zukunft und mögliche technische Entwicklungen – und zeigt, dass der Umgang mit künstlicher Intelligenz Verantwortung erfordert, aber auch viel Potenzial bietet. (spa) c't

Smart Glasses: Die neue Generation

Die smarten Brillen sind zurück – unauffälliger, leistungsfähiger und intelligenter denn je. Schafft die Produktkategorie jetzt den Durchbruch? Und was müssen Nutzer wissen, bevor sie sich eine zulegen?

Von **Nico Jurran**

Was sind Smart Glasses? Dass diese Frage nicht so einfach zu beantworten ist, zeigt schon ein kurzer Blick in die englische Wikipedia, die gleich drei, teilweise deutlich voneinander abweichende Definitionen liefert. Der kleinste gemeinsame Nenner ist meistens, dass es sich um Brillen handelt, in die elektronische Komponenten wie Sensoren, Kameras, Mikrofone, Lautsprecher oder Displays eingebaut sind. Bei einer so niedrigen Hürde zählen dazu aber auch Mixed-Reality-Headsets wie die Meta Quest oder Apples Vision Pro.

Diese Artikelreihe folgt einer strengeren Auslegung des Begriffs „smarte Brille“. Hier geht es um Modelle mit herkömmlichem Brillendesign, die über den Einsatz als Sehhilfe, Sonnenschutz oder modisches Accessoire hinaus zusätzliche Funktionen wie Medienwiedergabe, Videoaufzeichnung, Telefonie, Benachrichtigungen oder Navigation bieten, statt komplett auf immersive Erlebnisse ausgelegt zu sein. Anders ausgedrückt: Wenn bei diesen Geräten der Akku leer ist, lassen sie sich weiter als „dumme“ Brillen tragen. Daher behandeln wir in dieser Artikel-

strecke auch keine Videobrillen, mit denen man etwa im Flugzeug einen Film anschaut.

Die Brille, die wohl die meisten bislang mit dem Begriff Smart Glasses in Verbindung brachten, ist die 2012 erstmals präsentierte „Google Glass“. Sie erlebte zwar einen kurzen Hype, setzte sich in der Breite aber nicht durch. Doch wer danach meinte, die Produktkategorie sei ein für alle Mal tot, muss sich eines Besseren belehren lassen: So verkaufte sich die „Ray-Ban Meta“ von EssilorLuxottica und Meta (Test auf S. 82) seit der Markteinführung Ende 2024 weltweit bereits mehr als zwei Millionen Mal und ist aktuell das nach eigenen Angaben beliebteste Modell des Brillenkonzerns.

Eine neue Generation

Das blieb nicht ohne Folgen: Mittlerweile haben etliche Unternehmen smarte Brillen angekündigt, erste Konkurrenzmodelle zur Meta-Brille wie die „Even G1“ von Even Realities (siehe S. 86) sind sogar auch hierzulande bereits erhältlich. Auf der CES und der IFA präsentierten zudem eine Reihe chinesischer Firmen fast serienreife Smart Glasses. So will Rokid, das hierzulande bereits Videobrillen anbietet, sein Modell „Glasses“ im November fertig haben.

Google kehrt ebenfalls zurück: Gemeinsam mit Samsung arbeitet das Unternehmen an Smart Glasses mit dem speziell für dieses Einsatzgebiet entwickelten Betriebssystem Android XR (für „Extended Reality“). Sie wurden auf der diesjährigen Google I/O präsentiert und kommen wohl noch 2025 auf den Markt. Und auch Apple soll eine smarte Brille in der Entwicklung haben, deren Marktstart allerdings noch nicht für dieses Jahr erwartet wird.

Doch was ist bei den neuen Smart Glasses anders? Das fängt beim Erscheinungsbild an: Sah die Google Glass mit ihrem deutlich sichtbaren Display schon von Weitem so aus, wie man sich gemeinhin eine Datenbrille vorstellt, sind aktuelle Smart Glasses in der Regel nicht mehr von gewöhnlichen Brillen zu unterscheiden. Die komplette Elektronik samt Sensoren und Bedienelementen steckt hier unauffällig im Rahmen und in den Bügeln.

Trotzdem unterscheiden sich die aktuellen Smart Glasses stark voneinander. So übermitteln einige Modelle – wie die Ray-Ban Meta – Informationen ausschließlich per Audio an den Nutzer, während andere voll auf visuelle Darstellung setzen. Das kann etwa über ein winziges Display im Rahmen geschehen, zu dem man schielt. Der Trend geht aber klar in

Die smarte Ray-Ban Meta lässt sich auf den ersten Blick kaum von einer gewöhnlichen Brille unterscheiden – was zu ihrem Erfolg beigetragen haben dürfte.

Richtung Miniprojektoren, die digitale Inhalte von innen auf die Gläser werden und so ins Blickfeld des Nutzers einblenden. Das beherrscht die Even G1, wie wir auf Seite86 genauer erklären. Und auch Meta bietet das bei seiner Ray-Ban Meta Display, für die es hierzulande noch keinen Erscheinungstermin gibt.

Bitte mit KI

Gemeinsam haben die Smart Glasses der neuen Generation eine starke Verknüpfung mit dem Bereich der generativen künstlichen Intelligenz, weshalb auch von „KI-Brillen“ oder „AI Glasses“ die Rede ist. Man trägt also praktisch einen Assistenten auf der Nase, der jederzeit Fragen beantwortet – zum Beispiel solche nach Uhrzeit und Wetter. Die KI erlaubt daneben oft eine freihändige Bedienung der Funktionen, etwa über Sprachbefehle wie „Rufe Jan-Keno Janssen über WhatsApp an!“. Vor allem aber ermöglicht die KI Funktionen wie die Erkennung von Objekten und Personen oder Live-Übersetzungen von Sprache und Texten.

Damit ist aber auch klar, dass sich Smart Glasses von sogenannten Hör- oder Audiobrillen wie der Huawei Eyewear 2 (siehe Test in c’t 4/2024, S. 68) unterscheiden, die lediglich die Aufgaben eines Kopfhörers oder Headsets übernehmen und dafür per Bluetooth mit dem Handy gekoppelt werden. Und es eröffnet die Diskussion, wie smart Amazons (bislang nur in den USA erhältliche) Smart Glasses der „Echo Frames“-Reihe sind, die eine freihändige Verbindung zur „alten“ Alexa (ohne generative KI) auf dem Smartphone herstellen.

Bei den aktuellen und in näherer Zukunft erscheinenden Brillen laufen allerdings zumindest weitergehende KI-Funktionen über das gekoppelte Smartphone mit Begleit-App. Die aktuellen Wearable-Prozessoren sind zu schwach und zu ineffizient, um

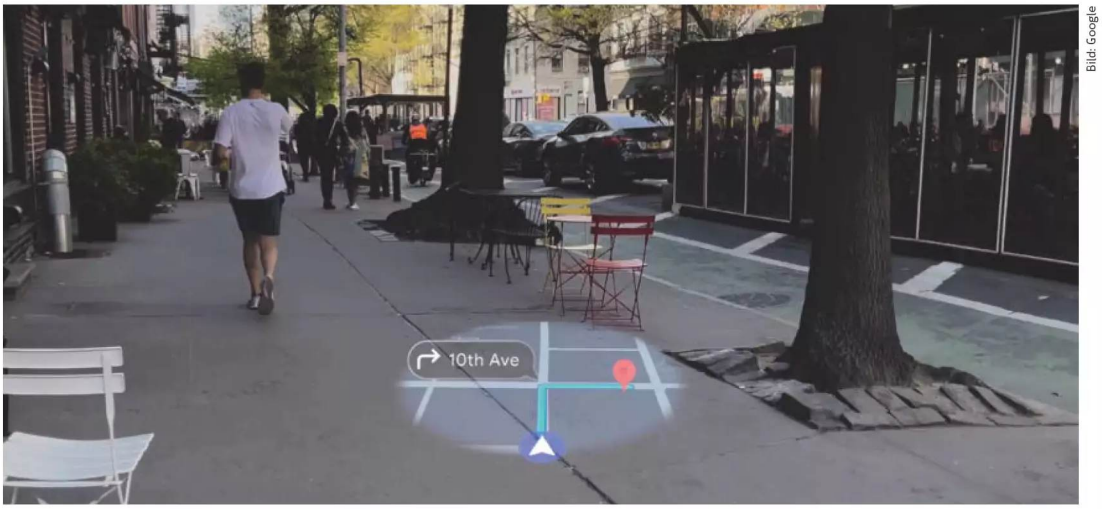


Bild: Google

Aus Googles Demovideo zur kommenden Brille mit Google XR: Anders als bei aktuellen Smart Glasses mit Projektion gibt es hier auch Einblendungen im unteren Bereich des Sehfeldes und in mehreren Farben.

diese Aufgabe zu übernehmen; zudem mangelt es ihnen an der nötigen Mobilfunkanbindung.

Und weil KI bei Smart Glasses eine große Rolle spielt, ist es auch wichtig, darauf hinzuweisen, dass bei der Ray-Ban Meta viele dieser Funktionen in EU-Ländern zunächst nicht freigeschaltet waren. Erst später lockerte Meta die Sperren – wobei hiesige Nutzer immer noch nicht auf alle Features zugreifen können, die US-Kunden zur Verfügung stehen. Das ist auch deshalb bemerkenswert, weil hinter den Einschränkungen weniger technische Gründe als Firmenpolitik standen und stehen. Letztlich sind die Kunden also den Launen des Herstellers ausgeliefert.

Einige Konkurrenten – wie Even Realities oder Rokid – haben darauf reagiert und werben nun damit, dass die Käufer bei ihren smarten Brillen selbst entscheiden können, welches KI-Modell zum Einsatz kommt. Doch schon bei der Even G1 zeigt sich auch die Kehrseite der Medaille: Da die Nutzung von KI Geld kostet, muss man bei Even Realities für Übersetzungen in bester Qualität Volumenkontingente kaufen.

Nicht ohne meine Brille

Dafür, dass sich Smart Glasses auf breiter Front durchsetzen, reicht es aber nicht, dass die Modellvielfalt in den kommenden Monaten signifikant steigt. Entscheidend ist, dass sich potenzielle Kunden davon überzeugen lassen, sich eine smarte Brille anzuschaffen. Das ist aber leichter gesagt als getan, wie die Erfahrung mit 3D-Fernsehen zeigt: Viele Menschen, die gewöhnlich keine Brille tragen, mochten eine solche nicht einmal für wenige Stunden auf der Nase haben.

Insofern ist es sicher kein Zufall, dass die Websites von Ray-Ban und Meta die KI-Brille vor allem als Sonnenbrille präsentieren: So will man sie offenbar auch Menschen ohne Sehschwäche schmackhaft machen. Auch die Assoziation von Sonne und Urlaub ist vom Hersteller wohl erwünscht. Schließlich liegt es nahe, die in der smarten Brille eingebaute Kamera für Schnappschüsse im Urlaub zu nutzen, während die Übersetzungsfunktion und die Objekterkennung bei der Orientierung vor Ort helfen.

Ein anderer Ansatz sind Fahrradbrillen, bei denen etwa Navigationshinweise in die Gläser projiziert werden. Dass diese Anzeigen konstruktionsbedingt bei der Even Realities G1 im oberen Bereich der Brillengläser erscheinen (siehe Seite 86), wäre hier durchaus vorteilhaft, da man beim Radfahren eher durch diesen Bereich der Brille schaut.

Wer im Alltag sowieso eine Brille trägt, hat üblicherweise eine niedrigere Hemmschwelle, sein gewöhnliches Modell gegen eine smarte Variante einzutauschen. Doch selbst diese Konstellation ist nicht immer glücklich. So können Brillenträger manche Smart Glasses nicht einfach als Sehhilfe nutzen. Das trifft vor allem auf Modelle zu, bei denen Informationen auf die Brillengläser projiziert werden, die dafür einen speziellen Schliff haben müssen. Wir widmen der Frage, was Brillenträger bei Smart Glas-

ses beachten müssen, daher einen eigenen Artikel ab Seite 90.

Achtung, Aufnahme!

Nicht außer Acht lassen sollte man die Rechtslage. So kann die unauffällige Integration von Kameras und Mikrofonen problematisch werden – Stichwort: unerwünschte und eventuell illegale Foto- und Videoaufnahmen. So verbietet § 201 a StGB etwa heimliche Aufnahmen in der Wohnung und in besonders geschützten Räumen wie Toiletten und Umkleiden. Und bei Audioaufzeichnungen (mit und ohne Video) kann man auch an § 201 StGB denken, der die Aufnahme des nichtöffentlich gesprochenen Wortes eines anderen auf einen Tonträger untersagt.

Nach Meinung des auf Datenschutz und IT-Recht spezialisierten Rechtsanwalts Stefan Hessels reicht aber die während der Aufnahme blinkende LED, damit die Ray-Ban Meta nicht unter § 8 des „Gesetzes zur Regelung des Datenschutzes und des Schutzes der Privatsphäre in der Telekommunikation und bei digitalen Diensten“ (TDDDG) fällt, der Spionagekameras in Alltagsgegenständen verbietet. Das unbemerkte Ausspionieren ist seiner Meinung nach so kaum möglich. Das gilt natürlich nicht, wenn man die LED manipuliert.

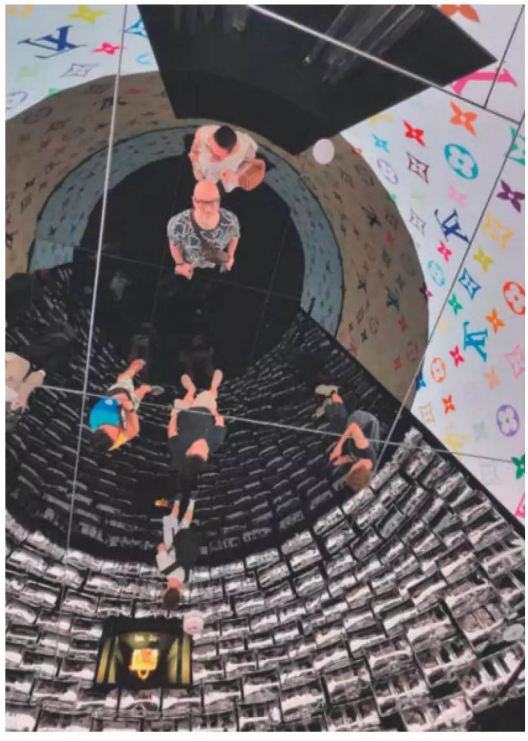
Inwieweit Aufnahmen mit Smart Glasses im öffentlichen Raum zu einem größeren Problem werden, wird sich zeigen müssen. Straßenszenen lassen sich mit der Kamera in der Brille genauso aufnehmen, wie es heute oft mit dem Handy gemacht wird. Denkbar ist, dass die jüngere Generation vor dem Hintergrund von TikTok & Co. öffentlichen Aufnahmen offener gegenüber steht. Aber ein gewisses Missbrauchspotential lässt sich kaum bestreiten.

Die Firma Solos hat sich als Lösung ausgedacht, die Elektronik und den Akku komplett in die Bügel zu verlagern und den Rahmen austauschbar zu machen. Neben einem Rahmen mit Kamera will die Firma dann auch solche ohne anbieten. So kann jeder Nutzer selbst entscheiden. Andere Hersteller wollen die Kamera lediglich zur Objekterkennung einsetzen, Foto- oder Filmaufnahmen aber nicht ermöglichen. Hier dürfte es im Zweifel dann aber immer noch Klärungsbedarf mit umstehenden Passanten geben. Die Even Realities G1 hat gar keine Kamera.

Wie geht es weiter?

Bei den smarten Brillen der neuen Generation drängt sich der Vergleich zu den ersten Smartwatches auf. Auch sie waren hinsichtlich des Funktionsumfangs beschränkt, boten oft nur sehr kurze Laufzeiten und waren sehr stark auf das Smartphone angewiesen. Das Hauptaugenmerk der Hersteller wird nach der ersten Welle an smarten Brillen daher auf leistungsfähigere und energieeffizientere Prozessoren liegen, damit kommende Modelle eine annehmbare Laufzeit bieten und mehr Features autark funktionieren.

Fest steht: Die neuen Smart Glasses markieren erst den Anfang einer Entwicklung, die den Alltag vieler Menschen nachhaltig verändern könnte – wenn Hersteller es schaffen, Technik, Tragekomfort, Datenschutz und praktische Funktionen überzeugend zu vereinen. Ob sich smarte Brillen dann tatsächlich durchsetzen, hängt also davon ab, ob sie den Sprung vom Gadget zum alltagstauglichen Begleiter schaffen. (nij) c't



Selfie mit der Ray-Ban Meta: Mit einer Kamera in einer Brille, die man sowieso trägt, ist die Aufnahme schneller gemacht, als man ein Handy aus der Tasche ziehen kann.

Smarte Brille: Ray-Ban Meta

Bevor eine Reihe anderer Smart Glasses auf den Markt kommt, hat Meta hierzulande die lang erwarteten KI-Funktionen seiner smarten Ray-Ban-Brille freigeschaltet. Doch es bleiben weiter Wünsche offen.

Von **Nico Jurran**

Als die smarte Brille Ray-Ban Meta auf den Markt kam, war nicht abzusehen, welchen Nerv der Brillenkonzern EssilorLuxottica und die Facebook-Mutter Meta damit treffen würden. Mittlerweile ist die zu Preisen ab 329 Euro erhältliche Meta-Brille das meistverkaufte Modell von Essilor-Luxottica, obwohl der Konzern auch noch Brillen von Marken wie Oakley, Persol und Prada vertreibt.

Der Erfolg ist umso bemerkenswerter, wenn man bedenkt, dass Meta die meisten KI-Funktionen in der EU erst Ende April dieses Jahres freischaltete. Die Ray-Ban Meta verkaufte sich hier also schon gut, als ihre Fähigkeiten noch eingeschränkt waren.

Das liegt auch daran, dass man bei ihrem Anblick nicht gleich an Smart Glasses denkt. Vor allem beim glänzend-schwarzen Wayfarer-Modell muss man genau hinschauen, um es nicht für das klassische Sonnenbrillenmodell zu halten. Dann sieht man, dass eine Kamera auf der linken und eine LED auf der rechten Ecke des Rahmens sitzen. Im Vergleich mit der gewöhnlichen Wayfarer fallen zudem die breiteren Bügel auf, in denen Akku und Elektronik

stecken. Mittlerweile gibt es auch eine Oakley-Variante der Meta-Brille, die auf den Namen Oakley Meta HSTN hört, etwas teurer ist und sich in einigen Details von der Ray-Ban Meta unterscheidet – dazu gleich mehr.

Was ging, was geht?

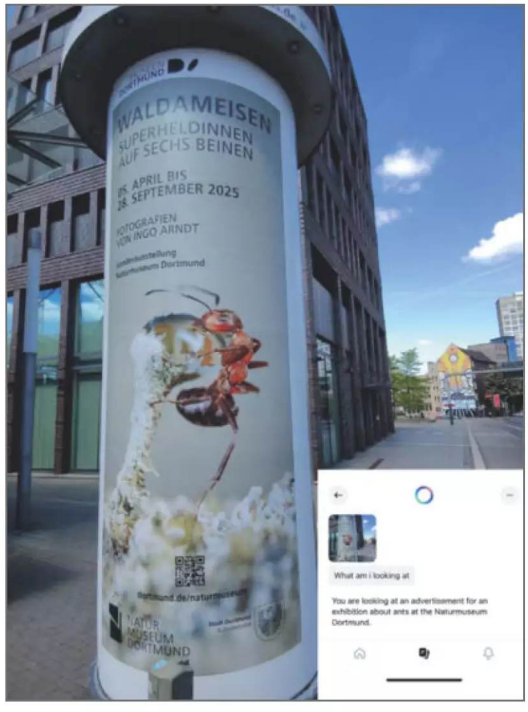
Fünf Mikrofone und zwei Lautsprecher im Rahmen und in den Bügeln erlauben es bei der Ray-Ban Meta, Musik und Podcasts abzuspielen, zu telefonieren, Timer zu stellen und Sprachnachrichten zu verschicken. Wohlgemerkt läuft dies alles über das per Bluetooth verbundene Smartphone; die Brille selbst hat weder Audioplayer noch Mobilfunk. Die Audioqualität ist recht beachtlich, auch wenn sie an gute Ohrhörer nicht heranreicht. Bei Telefonaten lobten die Gesprächspartner die gute Verständlichkeit.

Die Außenseite des rechten Bügels ist (deaktivierbar) berührungsempfindlich, sodass man mit Wischgesten die Lautstärke regeln und die Wiedergabe durch Antippen starten und stoppen kann. Alternativ ruft man die Funktionen per Sprache mit „Hey Meta“ auf. Da sich die Brille über die Begleit-App „Meta AI“ (vormals „Meta View“) mit einigen Diensten und Plattformen verknüpfen lässt, kann man nach dem Muster etwa auch direkte Musik von Apple Music abrufen oder Telefonate über WhatsApp führen. Schließlich informiert die Meta-Brille über eintreffende Nachrichten und liest diese auf Befehl vor.

Spätestens an dieser Stelle ist wichtig, dass die Brille (wie die App) nicht nur Englisch, Spanisch, Italienisch und Französisch, sondern mittlerweile auch Deutsch beherrscht. Las sie vorher deutsche Mitteilungen vor, verstand man oft kein Wort. Leider hat man mit Deutsch als Systemsprache aber nicht Zugriff auf alle KI-Funktionen. Außen vor bleibt dann etwa für die Live-Übersetzungsfunktion, die mit Englisch als Zielsprache hingegen zur Verfügung steht.

Die Übersetzung läuft dabei komplett auf der Brille, weshalb man für die zu übersetzende Sprache zunächst das passende Datenpaket auf die Ray-Ban Meta herunterlädt. Nach dem Start bekommt man satzweise ins Ohr gesprochen, was das Gegenüber gesagt hat – und auch parallel eine Transkription auf dem Smartphone. Das funktionierte im Test gut.

Schließlich kann man der Meta-KI beliebige Fragen stellen, wie man dies von ChatGPT & Co. kennt – inklusive direkter Nachfragen, ohne erneutes Hotword. Auf diesem Weg kann man auch die Zeit erfragen oder sich das Wetter vorhersagen lassen.



Test der Objekterkennung anhand einer Werbung: Die Meta-KI erkannte nicht einfach die Ameise, sondern teilte uns über die Brille konkret mit, um welche naturkundliche Ausstellung zu diesen Tieren es sich handelte.

Achtung, Aufnahme!

Das Highlight aber bleibt, mit der Brille Fotos und Videos mit einer Länge von bis zu drei Minuten aufnehmen zu können – passend für Social Media im Hochkantformat. Ausgelöst werden die Aufnahmen über einen Knopf auf dem rechten Bügel oder per Sprachbefehl, auch ohne Handy in der Nähe. Den dabei auftretenden Datenschutzbedenken begegnet Meta, indem es die angesprochene LED bei den Aufnahmen blinken beziehungsweise dauerhaft leuchten lässt. Deckt man sie ab, verweigert die Brille den Dienst.

Bei Fotos ist eine solche Manipulation aber nicht mal nötig: Kaum jemand bemerkt das kurze Blinken. Wer daher meint, mit der Meta-Brille ein Spannerwerkzeug zu erwerben, wird aber wohl enttäuscht sein. Die integrierte Kamera ist sehr weitwinklig,

Ray-Ban Meta Wayfarer	
Smarte Brille mit Kamera	
Hersteller, URL	EssilorLuxottica, ray-ban.com
Prozessor, Speicher	Qualcomm Snapdragon AR1 Gen 1, 32 GByte (EMMC)
AV-System	Kamera (Videos: 1440 × 1920 Pixel bei 30 fps, Fotos: 3024 × 4032 Pixel), 2 Mikrofone, 5 Lautsprecher (Open-Ear-Array)
Konnektivität	Wi-Fi 6, Bluetooth 5.3
Systemanf.	Smartphone mit Android 14.4+ oder iOS 10.0+
Maße (H × B × T) / Gewicht / Schutzart	4,6 cm × 14 cm × 15,3 cm / 49 g / IPX4
Preis	ab 329 €

weshalb Personen, die nicht unmittelbar vor einem stehen, eher schmückendes Beiwerk sind. Mit ihrer Auflösung von 12 Megapixeln bei Fotos und 1440 × 1920 Pixeln bei Videos und ohne Bildstabilisierung oder irgendwelchen Modi lieferte die Kamera durchaus gelungene Schnappschüsse, reichte an aktuelle Smartphones aber bei Weitem nicht heran. Eine jüngst erschienene 2. Generation bietet einen höherwertigen Sensor, der Videos in 3K-Auflösung aufnimmt.

Seit dem KI-Update lässt sich die Kamera auch zur Erkennung von Objekten und bekannten Persönlichkeiten nutzen und um Texte übersetzen zu lassen. Dafür knipst die Brille auf Sprachbefehl ein Bild und lädt es über die Meta-AI-App in die Meta-Cloud zur Analyse hoch. Das Ergebnis teilte die Brille per Sprache mit. Die Erkennung funktionierte im Test erstaunlich gut. Blinde und sehbehinderte Menschen können über das Netzwerk von „Be My Eyes“ Unterstützung bei alltäglichen Aufgaben durch sehende Freiwillige erhalten. Dafür teilt die Brille ihr Kamerabild und ermöglicht die Kommunikation zwischen Nutzer und Helfer.

Modellauswahl

Neben der getesteten Wayfarer gibt es mit Skyler ein weiteres Grundmodelle der Meta-Brille, zudem bringt Ray-Ban immer mal limitierte Editionen heraus. Die Technik ist stets identisch. Das Design wirkt ansprechend, auch wenn Verarbeitungsqualität und haptisches Erlebnis hinter den „dummen“ Ray-Bans zurückbleiben. Aufgrund der starren Bügel sollte man vor dem Kauf genau prüfen, ob die Brille gut sitzt.

Wer möchte, kann ab Werk Korrekturgläser bestellen oder solche nachträglich beim Optiker einsetzen lassen. Da dabei einige Punkte zu beachten sind, widmen wir diesem Thema einen eigenen Beitrag ab Seite 90. Ray-Ban liefert die Meta mit einem Etui aus festem Kunststoff aus, das zugleich als Ladestation dient und die Brille über seinen integrierten Akku achtmal komplett aufladen kann. Der proprietäre Ladeanschluss am Steg der Meta-Brille geht eine feste Verbindung mit dem Gegenpart im Etui ein, weshalb man sie manchmal nur recht schwer wieder herausbekommt.

Leider ist der fest integrierte 154-mAh-Akku der Brille selbst bei moderater Nutzung schon nach vier Stunden am Ende. Bei der 2. Generation wurde die Akkuleistung verbessert, so dass sie nun laut Meta im Normalbetrieb bis zu acht Stunden durchhält.

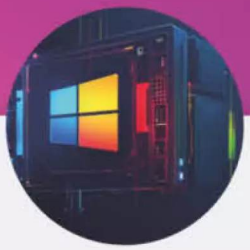
Fazit

Im ersten Augenblick wirkt die Ray-Ban Meta wie eine Spielerei. Doch die smarte Brille wird schneller zum Alltagsgegenstand, als man denkt – weil sich Musik und Podcasts darüber anhören lassen, ohne ihren Träger komplett von der Umwelt zu isolieren. Und weil sie auf Zuruf freihändige Telefonate ermöglicht, ohne dass man erst die Ohrhörer einstöpseln muss. Im Urlaub fängt man mit der Brille, auch dank Sprachsteuerung, spontaner und natürlicher Fotos und Videos von Straßenszenen ein, als wenn man erst sein Handy herauskramt.

Da die KI nun auch Deutsch als Systemsprache beherrscht, kann die Brille auch endlich eintreffende Mitteilungen vernünftig vorlesen. Bedauerlich ist, dass sie dann bei KI-Funktionen wie dem Übersetzungsfeature weiter hinterherhinkt. Immerhin besteht hier die Hoffnung, dass Meta der Brille auch dies noch per Update beibringt. Kein Firmware-Update kann hingegen die sehr mäßige Laufzeit beheben, die umso stärker ins Gewicht fällt, je mehr die Brille zum täglichen Begleiter wird. (nij) c't

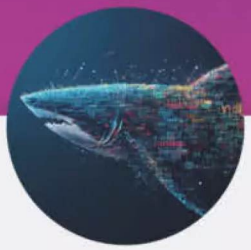
Aktuelle Workshops

iX Workshop



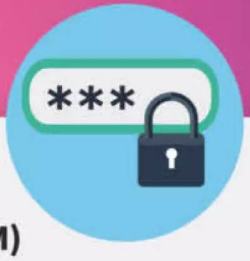
19. – 20. November

Windows Server absichern und härten



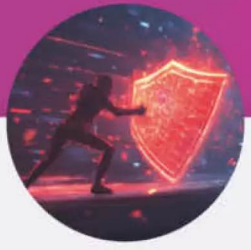
24. – 26. November

Netzwerkanalyse und Fehlersuche mit Wireshark



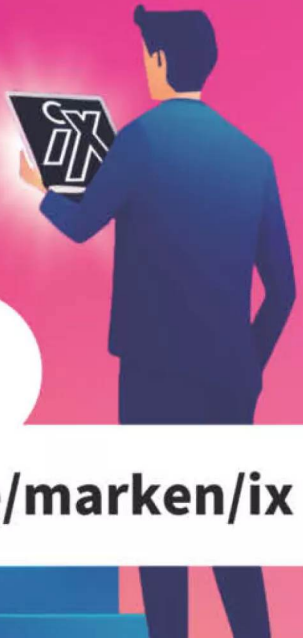
25. – 27. November

Identity & Access Management (IDM/IAM) und SSO mit Keycloak



26. – 28. November

Active Directory Hardening
Vom Audit zur sicheren Umgebung



JETZT SKILLS UPGRADEN

auf **heise-academy.de/marken/ix**

Smarte Brille: Even Realities G1

Even Realities' smarte Brille Even G1 lässt Informationen vom Handy vor den Augen des Nutzers schweben. Das ermöglicht faszinierende Features wie einen virtuellen Teleprompter, die Umsetzung birgt aber auch Schwächen.

Von **Nico Jurran**

Dass das Hongkonger Unternehmen Even Realities mit seiner smarten Brille „Even G1" nicht einfach die erfolgreiche Ray-Ban Meta (siehe Seite 86) nachgebaut hat, zeigt sich schon an den Bügeln. Sie sind hier nicht durchgehend und bestehen nicht aus starrem Kunststoff, sondern aus einer mit Silikon überzogenen, flexiblen Titanlegierung und haben nur an den Enden kleine Blöcke, in denen Elektronik, Akkus und Touchfelder stecken. Im Test empfanden viele – über eine breite Spanne an Kopfformen hinweg – das Tragegefühl als angenehm.

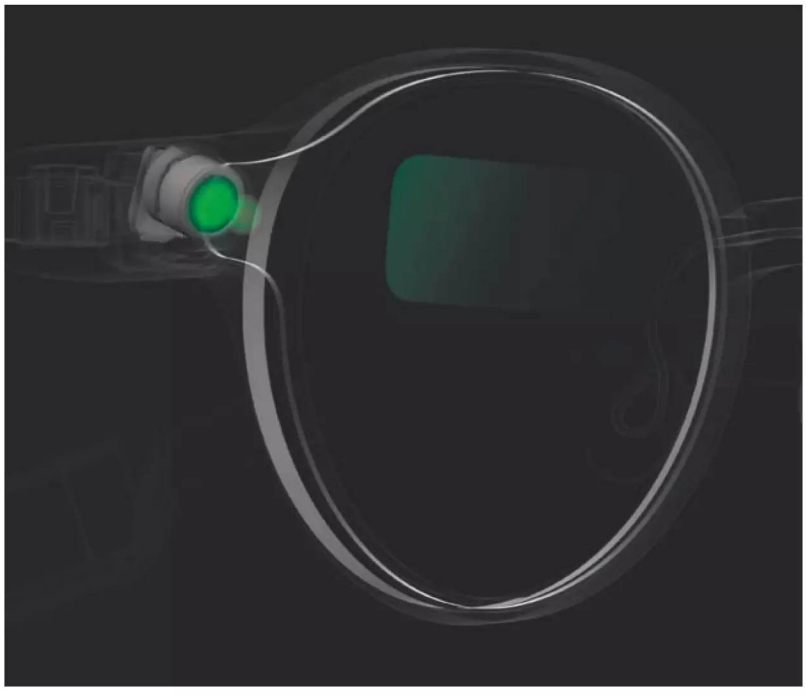
Der eigentliche Aha-Effekt tritt aber ein, wenn man die 699 Euro teure Brille aufsetzt und die Begleit-App auf dem (über zwei Verbindungen) per Bluetooth gekoppelten Smartphone startet: Nun scheinen vom Handy übertragene Texte und einfache Grafiken vor einem mittig im Raum in einer Entfernung von rund zwei Metern zu schweben. Zwar hat die Anzeige nur eine Auflösung von 640 × 200 Pixel und erscheint ausschließlich in Grün, dennoch ist der Effekt beeindruckend.

Möglich machen dies zwei Micro-LED-Projektoren in den Ecken des Rahmens der G1, die auf die Innenseiten der beiden Brillengläser projizieren. Passend dazu stecken in der Brille „Waveguide"-Gläser, die das vom optischen System ausgestrahlte Bild zum Auge des Benutzers leiten und gleichzeitig selbst

klar und transparent sind. Jedes Glas enthält innen eine Schicht mit einer rund 2,8 Zentimeter × 1,3 Zentimeter großen, etwas matteren Projektionsfläche. Schaut man durch diesen Bereich, erkennt man an diesen Stellen einen minimalen Helligkeitsunterschied, der im Alltag aber nicht stört.

Menschen mit Sehschwäche sehen ohne passende Gläser statt des Textes oft nur einen grünen Blob. Doch Optiker können Korrekturgläser nicht einfach mit der Waveguide-Technik ausstatten. Daher bietet Even Realities die G1 für einen pauschalen Aufpreis von 150 Euro mit passenden Einstärkengläsern an. Getönte Gläser gibt es nicht, für 100 Euro aber einen Sonnenbrillen-Clip, den man vorne am Rahmen einhakt. Damit lässt sich die Anzeige auch bei strahlendem Sonnenschein gut ablesen. Ohne den Clip bekommt man hingegen schon in Innenräumen Probleme, den Text zu lesen, wenn man etwa an einem hellen Tag zum Fenster guckt.

Bild: Even Realities



Die optische Mikro-LED-Engine der Even G1 überträgt die Inhalte auf ein Waveguide-Glas. Der Miniprojektor im Rahmen ist mit bloßem Auge nicht zu erkennen.

Souffleuse

Es drängt sich auf, die Even G1 als virtuellen Spickzettel zu nutzen. Und tatsächlich lässt sich mit der Funktion „Teleprompter“ ein beliebiges Skript als Text- oder Word-Datei in die Even-App laden, das die Brille dann anzeigt. Besonders gelungen: Die G1 erfasst über zwei Mikrofone, was der Nutzer sagt, und kann diese Eingaben mittels KI mit dem Text abgleichen. So scrollt das Skript zum richtigen Zeitpunkt automatisch weiter – auch wenn man zwischendurch eine Pause macht oder etwas von der Vorlage abweicht.

Mit korrekt angepassten Gläsern ist die Schrift so scharf, dass das Ablesen des Textes vielen leichter fällt als von einem kleinen Teleprompter. Das Publikum bekommt von der Projektion meist nichts mit, sondern bemerkt bestenfalls, dass der Sprecher ein wenig nach oben schaut. Das ist dem Umstand geschuldet, dass sich die Projektoren auf der Höhe der Bügel befinden und somit auch die Projektionsflächen im oberen Drittel der Gläser. Die Anzeige lässt sich zwar etwas verschieben, am jeweiligen Maximum verblasst die Schrift aber am oberen beziehungsweise unteren Rand.

Mit der G1 kann man sich zudem im Urlaub das Wörterbuch sparen: Die Brille erfasst auf Wunsch über ihre Mikrofone oder das gekoppelte Handy, was das Gegenüber in einer von 24 Fremdsprachen sagt, und zeigt die Übersetzung auf Deutsch oder einer von 19 weiteren Sprachen an. Das funktioniert nicht immer fehlerfrei, der Sinn der Aussagen kommt aber meist gut rüber. Die Funktionen „Transkribieren“ und „Notizen“ machen wiederum ohne Übersetzung sichtbar, was der Nutzer beziehungsweise seine Gesprächspartner sagen.

Leider bleibt bei der Übersetzungsfunktion ein Beigeschmack. Denn Even Realities bietet eine höhere Qualität nur gegen Bezahlung an: Eine Stunde kostet 5,99 Euro, fünf Stunden 24,99 Euro. Kostenlos ist diese Pro-Version nur, wenn man die Brille vor Ende August 2024 geordert hat. Laut Even Realities können aber alle Käufer die kostenlose Version mit niedrigerer Übersetzungsqualität unbegrenzt nutzen.

Navi auf der Nase

Naheliegend ist auch die Nutzung der G1 als Head-up-Display für Navigationshinweise. Da die Daten für die G1 aufbereitet werden müssen, lassen sich dafür nicht Google Maps & Co. verwenden, sondern

Even Realities Even G1 B	
Smarte Brille mit Projektion	
Hersteller, URL	Even Realities, evenrealities.com
Prozessor, Speicher	Snapdragon XR2, k.A.
AV-System	2 Micro-LED-Projektoren (640 × 200 Pixel bei 20 fps, einfarbig, lt. Herst. 1000 Nits Helligkeit), 2 Mikrofone
Konnektivität	Bluetooth 5.2
Systemanf.	Smartphone mit Android 14.4+ oder iOS 16.0+
Maße (H × B × T) / Gewicht / Schutzart	4,3 cm × 13,4 cm × 16,1 cm / 45 g / k. A. (geschützt vor Spritzwasser und leichtem Regen)
Preis	699 €, 150 € Aufpreis für Einstärkengläser, 100 € für Sonnenbrillen-Clip

nur die Even-App. Diese bietet Fußgänger- und Radfahrernavigation, ist kartenmäßig aber schlecht aufgestellt: Auf Testfahrten schickte uns die G1 an Autostraßen entlang, obwohl Rad- und Fußwege schneller zum Ziel geführt hätten. Zu allem Überfluss baute die App Umwege ein, wodurch sich eine Radstrecke von den üblichen 18 auf 26 Minuten verlängerte.

Da Even Realities voll auf die visuelle Vermittlung von Informationen setzt, sind in in der G1 keine Lautsprecher eingebaut. Damit fällt eine Wiedergabe etwa von Musik oder das Telefonieren über die Brille komplett flach. Wer möchte, kann einblenden lassen, wenn ein Telefonat am gekoppelten Handy ankommt, oder sich eine dort eingetroffene WhatsApp-Nachricht anzeigen lassen.

Daneben kann man der KI mündlich Fragen stellen und bekommt darauf Antworten in Textform – leider ohne die Möglichkeit direkter Nachfragen. Die Besonderheit: Die G1 läuft im Unterschied zur Meta-Brille nicht nur mit dem GenAI-Modell des Herstellers, sondern alternativ auch mit Perplexity (voreingestellt) und ChatGPT.

Im Alltag beschränkt sich die Interaktion mit der Brille ansonsten auf ein Dashboard, das eingeblendet wird, wenn man den Kopf hebt. Dieses ist in Grenzen konfigurierbar, sodass man sich neben Datum, Uhrzeit und Terminen etwa Notizen, Aktienkurse oder aktuelle Schlagzeilen anzeigen lassen kann.

Modellauswahl

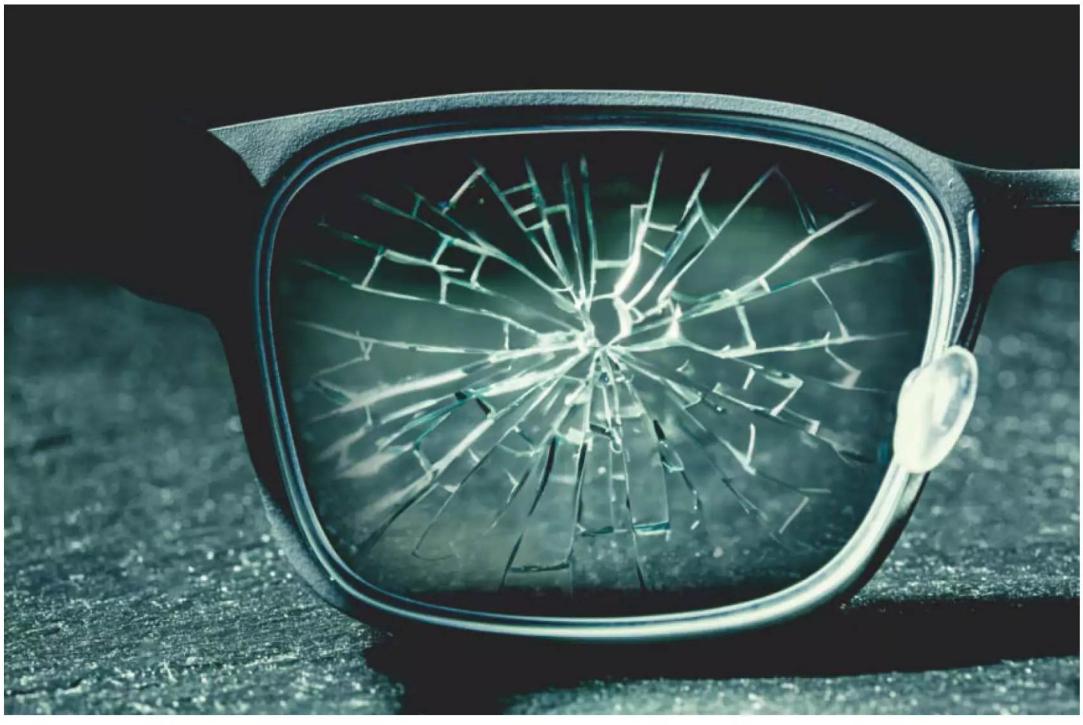
Even Realities bietet die Even G1 mit zwei verschiedenen Rahmen an: die getestete eckige Variante (Modell B) und eine runde Fassung (Modell A). Technische Unterschiede gibt es zwischen beiden nicht.

Der fest integrierte 160-mAh-Akku hält laut Hersteller bei regelmäßiger Nutzung anderthalb Tage durch. Das kommt nach unserer Erfahrung hin – wobei aber auch nur wenige Nutzer durchgehend die Projektion aktivieren dürften. Wie die Meta-Brille kommt die Even G1 mit einem Etui, das zugleich als Ladestation dient. Dieses hängt die Ray-Ban-Variante schon deshalb um Längen ab, weil die Aufladung des Brillenakkus hier drahtlos läuft und man die G1 folglich nur hineinlegen muss. Der im Etui integrierte Akku reicht für zweieinhalb weitere Runden.

Fazit

Even Realities ist mit der Even G1 zweifellos ein faszinierendes Gadget gelungen. Und wer häufig Vorträge hält, oft Videos produziert oder internationale Messen besucht, kann sich leicht vorstellen, diese Brille im Einsatz zu haben – auch wenn für eine Konversation mit einem ausländischen Gesprächspartner dieser ebenfalls Smart Glasses tragen müsste.

Genau hier liegt auch die Achillesferse der Even G1: So sehr sie bei solchen speziellen Einsätzen ihre Stärke ausspielt, so stark fällt sie im Alltag ab – auch gegen die Ray-Ban Meta. So dürften viele auf das Dashboard und die Notizfunktion verzichten können. Wer aber schon mal die Ray-Ban Meta getragen hat, vermisst bald die Möglichkeit, einfach über die Brille Musik zu hören oder zu telefonieren. Daher wird die Even-Brille es schwer gegen kommende Modelle wie die Rokid Glasses haben, die beide Welten vereinen. (nij) c't



Smart Glasses für Brillenträger

Eigentlich sollten Smart Glasses doch wie gemacht sein für Menschen, die wegen einer Fehlsichtigkeit sowieso eine Brille tragen. Doch in der Praxis erwarten gerade sie einige zusätzliche Hürden.

Von **Nico Jurran**

Für Sie ist eine Brille Sehhilfe und nicht nur modisches Accessoire? Sie überlegen daher, Ihr gewöhnliches Gestell gegen eine smarte Variante auszutauschen? Dann könnten Sie dem Irrglauben erliegen, dass für Sie der Umstieg auf Smart Glasses einfacher wird als für Menschen, die im Alltag keine Brille tragen. Tatsächlich erwarten Sie zusätzliche Hürden – beim Kauf, im Reparaturfall und im täglichen Gebrauch.

Der erste Punkt dürfte schon deshalb einige verwundern, weil man im offiziellen Online-Shop die Ray-Ban Meta mit Einstärken- und Gleitsichtgläsern bestellen kann, auch in getönten oder selbsttönenden Varianten, entspiegelt und auf Wunsch extra

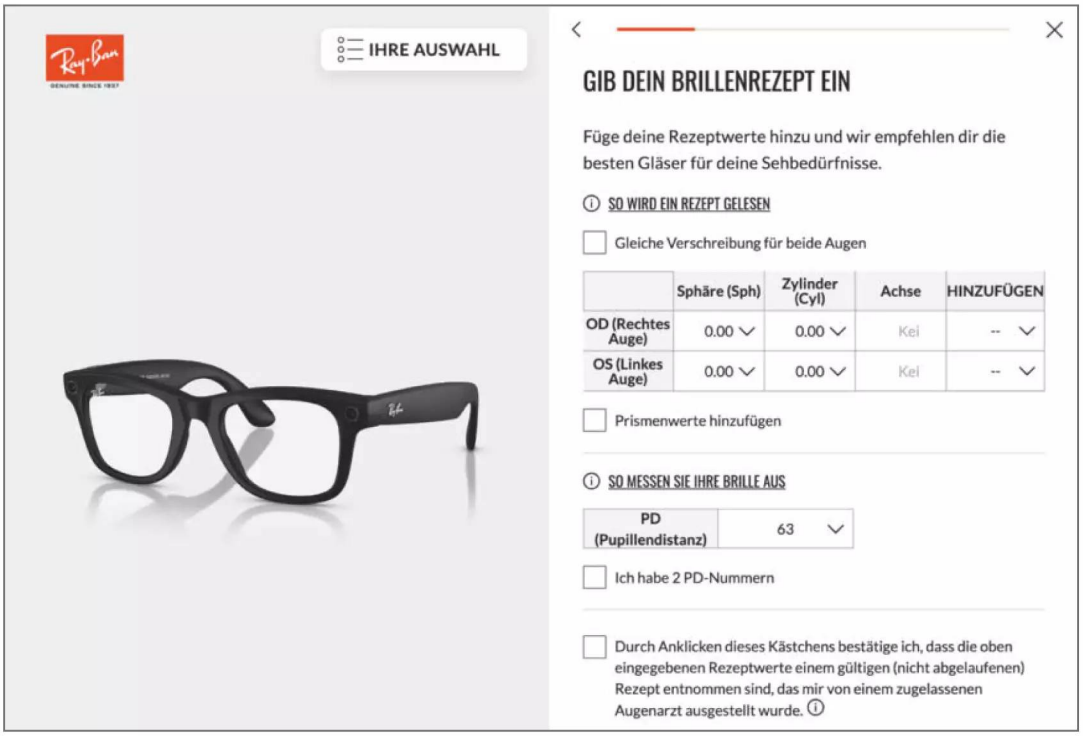
dünn. Dabei muss man ein aktuelles Rezept von einem zugelassenen Augenarzt hochladen beziehungsweise bestätigen, dass die eingegebenen Werte von einem solchen stammen.

Damit die Brille ihren Zweck als Sehhilfe erfüllt, müssen jedoch nicht nur die Gläser passend geschliffen sein, sondern auch an der richtigen Stelle sitzen, die Brille also der Kopfform angepasst sein. Denn während Optiker bei gewöhnlichen Brillen mit einem solch massivem Gestell noch Anpassungen vornehmen können, indem sie die Bügel leicht erhitzen und formen, ist dies bei der Ray-Ban Meta nicht möglich. Das Erhitzen könnte die Elektronik in den Bügeln beschädigen und fällt somit flach. Was nicht passt, passt nicht.

Sitzt die online georderte Meta-Brille also schlecht, bleibt somit oft nur die Rücksendung. Zwar besteht laut Meta auch bei Modellen mit Korrekturgläsern das 30-tägige Rückgaberecht, bis zur Rückzahlung des gegenüber den Modellen mit Fensterglas teilweise erheblich höheren Kaufpreises vergeht aber einige Zeit.

Beim Optiker

Daher könnte man überlegen, die Ray-Ban Meta mit Gläsern ohne Sehstärke online zu erwerben und erst bei Gefallen beim Optiker vor Ort mit Korrekturgläsern versehen zu lassen. Laut Ray-Bans Website bleibt die Herstellergarantie „dabei jedoch nur dann erhalten, wenn die Korrekturgläser von einem zertifizierten Händler eingesetzt werden“. Wie absurd diese Aussage für europäische Kunden ist, zeigt ein Blick auf das dazugehörige Onlinedokument (siehe ct.de/wdwg). Bis zum Redaktionsschluss waren darin passende Händler nur in den USA und Kanada zu



Die Meta-Brille lässt sich bei Ray-Ban gleich mit Korrekturgläsern bestellen – auch in einer Gleitsichtversion.

finden. Mancher dürfte die Gläser daher beim Optiker seines Vertrauens einsetzen lassen und Ray-Ban dies im Zweifelsfall verschweigen.

Nach Ablauf der Garantie bringt der eine oder andere seine Ray-Ban Meta aber wohl eh zum Optiker, um neue Gläser einsetzen zu lassen. Schließlich ändern sich bei Menschen mit Sehschwäche die Werte mit der Zeit, zudem sind Kratzer bei häufigem Gebrauch kaum zu vermeiden. Und nach unserem Kenntnisstand bietet Ray-Ban keine Gläser separat an.

Da es so wichtig ist, dass die Meta-Brille auf Anhieb gut sitzt, könnte man überlegen, die Brille gleich beim Optiker zu kaufen und dort mit Korrekturgläsern ausstatten zu lassen. Das ist jedoch oft einfacher gesagt als getan: Die Auswahl an unterschiedlichen Rahmen, Größen und Farben führt nicht selten dazu, dass gefragte Meta-Modelle vergriffen sind – teilweise für Monate. Besonders blöd ist, wenn ein Optiker zwar das gewünschte Modell in der richtigen Größe vorrätig hat, aber nur mit selbsttönenden Gläsern ohne Sehstärke. Greift man hier zu, bezahlt man einen Aufpreis für Gläser, die nie zum Einsatz kommen.

Projektionsbrillen

Trotz allem ist die Meta-Brille noch vergleichsweise unproblematisch, da sie sich nachträglich mit jedem Korrekturglas versehen lässt. Kniffliger wird es bei Brillen, die Informationen mittels Projektion auf den Innenseiten der Gläser anzeigen. Dafür stecken in der Even G1 von Even Realities etwa spezielle „Waveguide“-Gläser, die praktisch aus zwei Schichten bestehen – auch in der Standardversion ohne Sehstärke.

Even Realities bietet die Even G1 gegen einen Aufpreis von 150 Euro mit Einstärkengläsern in der Waveguide-Variante an. Für Gleitsichtgläser verweist der Hersteller hingegen auf Partner-Optiker. Mit dem Optiker sollte man auf jeden Fall besprechen, ob das Sichtfeld nicht zu klein wird, wenn man auch noch die für die Projektion nötige Fläche abrechnet. Getönte Gläser gibt es bei Even G1 nicht, sondern nur einen Sonnenbrillen-Clip (für 100 Euro), den man vor die Brille klemmt.

Der Hersteller Rokid bietet seine kommende KI-Brille mit Projektion einen Clip mit Korrekturgläsern

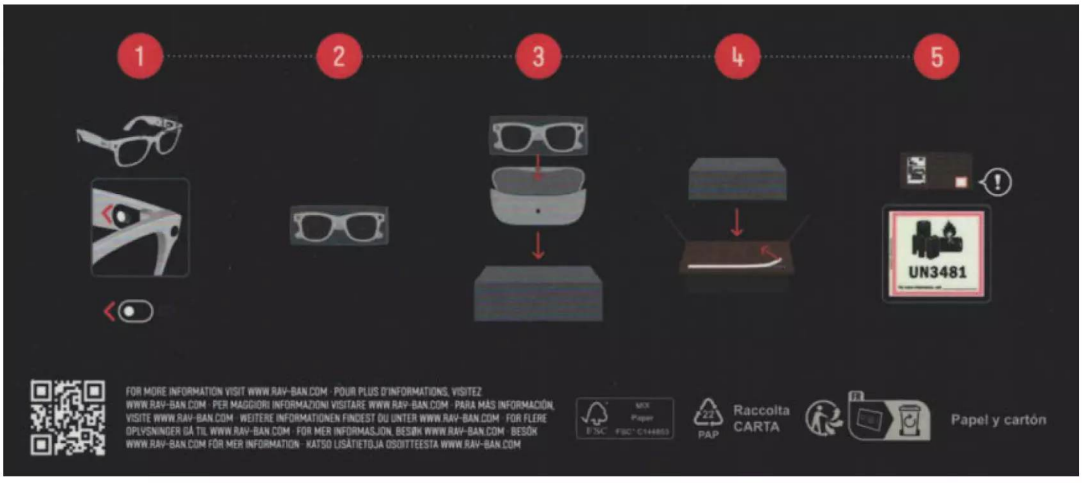
Im Schadensfall

Was macht man, wenn die Elektronik einer smarten Brille kaputtgeht? Mit dieser Frage musste ich mich notgedrungen auseinandersetzen, als meine privat erworbene Ray-Ban Meta plötzlich zickte: Mal ließ sie sich nicht starten, mal fielen Mikrofon und Kamera aus. Die Akkuanzeige wies derweil „-1“ oder Werte im Milliardenbereich aus.

Erste Erkenntnis: Wer die Brille in Metas Onlineshop erworben hat, kann sich die Kontaktaufnahme dorthin sparen. Am Ende verweist Meta eh auf den Service von EssilorLuxottica, der Firma hinter der Marke Ray-Ban. Auf deren Website füllt man ein Formular mit der Fehlerbeschreibung aus und erhält nach erfolgreicher Prüfung den Link zu einem weiteren Formular. Dort trägt man seine Adressdaten ein und vereinbart anschließend telefonisch über eine 0800-Nummer mit dem deutschsprachigen Support die Rücksendung.

Denn die Brille lässt sich nicht in einem beliebigen Paket retournieren. Stattdessen schickt EssilorLuxottica eine Umverpackung per UPS von einem Firmensitz in Italien – und in ebendieser wird die Brille dann vom Kurierdienst auch wieder dorthin zurücktransportiert. Die Kosten dafür musste ich in meinem Fall, der in die zweijährige Garantie fiel, nicht tragen. Allerdings nahm die Aktion etwas Zeit in Anspruch: Am Ende dauerte es 18 Tage von der Schadensmeldung bis zur Ankunft der Rücksendung. Aber immerhin erhielt ich eine komplett neue Brille.

Bild: Ray-Ban



Bastelstunde: Im Reparaturfall muss man für die Einsendung der Meta-Brille bei Ray-Ban ein Rücksendepaket anfordern.

anbieten, der zwischen Augen und Brillengläser kommt. Hier wird man schauen müssen, ob die Korrekturgläser das Sichtfeld nicht merklich einschränken. Mit dem Prototyp eines anderen Brillenmodells hatten wir zudem das Problem, dass die Wimpern immer wieder unangenehm gegen die Gläser eines solchen Clips schlugen.

Reparaturfall

Wenn die Elektronik kaputtgeht, macht es keinen Unterschied, ob die Brille über einen Onlineshop oder beim Optiker vor Ort gekauft wurde. In diesem Fall muss sie so oder so an den Hersteller zur Reparatur oder zum Austausch zurück. Mehr dazu im Kasten „Im Schadensfall“.

Wichtig: Bei einer Rücksendung müssen stets die ursprünglichen Gläser in der Ray-Ban Meta stecken – welche das sind, hat der Hersteller bei der Produktion erfasst und es ist auch in der Meta-AI-App nachzulesen. Man sollte diese also gut aufbewahren, wenn man beim Optiker nachträglich Korrekturgläser hat einsetzen lassen.

Apropos Rücksendung: Wer eine defekte Meta-Brille mit Korrekturgläsern retourniert, erhält von Ray-Ban vor der Einsendung den Hinweis, dass das Unternehmen für etwaige Versandschäden nicht haftet.

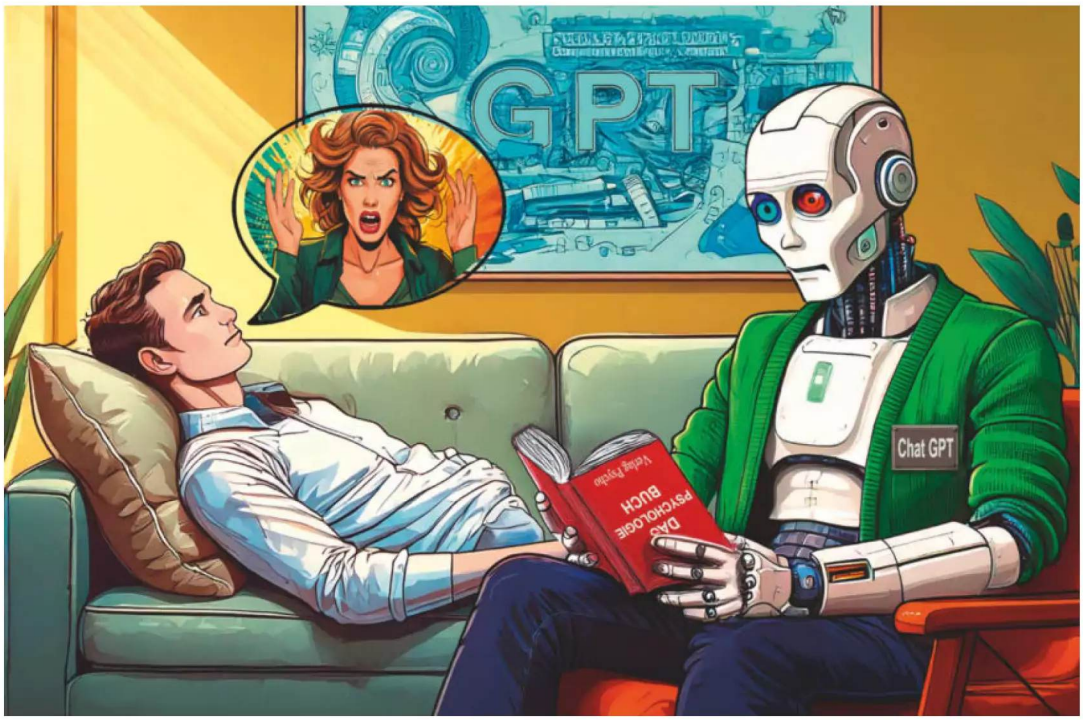
Meta-Dokument zur Ray-Ban-Brille:
ct.de/wdwg

Vorsicht im Alltag

Wer eine Ray-Ban Meta mit Korrekturgläsern als Alltagsbrille einsetzen möchte, sollte sich der geringen Akkulaufzeit des Modells bewusst sein: Der Hersteller selbst gibt nur vier Stunden am Stück mit einer Ladung an. Wer die Brille nicht für den Rest des Tages als dummes Gestell auf der Nase sitzen haben will, muss folglich eine Ersatzbrille mitführen für die Zeit, in der die Meta wieder aufgeladen wird. Spätestens an dieser Stelle sollte klar sein, dass es keine gute Idee ist, auf eine smarte Brille als einzige Sehhilfe zu setzen. Im Reparaturfall steht man dann erst einmal ohne Brille da.

Zum Abschluss ein allgemeiner Hinweis: Mit dem täglichen Einsatz von Smart Glasses steigt die Gefahr, dass man das smarte Modell irgendwann wie eine gewöhnliche Brille behandelt. Mittlerweile häufen sich Berichte von Nutzern, die etwa ihre Ray-Ban Meta „aus Reflex“ ins Spülwasser getaucht haben, um wie bei ihren anderen Brillen deren Gläser zu reinigen. Denn wirklich wasserdicht sind smarte Brillen in der Regel nicht.

Auch Even Realities spricht nur davon, dass ihre Even G1 geschützt sei vor Spritzwasser und leichtem Regen. Zumindest aktuell sind Smart Glasses hinsichtlich des Handlings also eher empfindliches Gadget als robuster Alltagsgegenstand. (nij) c't



(Bild: Jessica Nachtigall/KI/heise medien)

Menschen in Not bei ChatGPT in Therapie

Längst haben Menschen angefangen, mit KI-Chatbots über Sorgen zu reden, woraus sich in Krisen fast so etwas wie ein Therapiegespräch entwickeln kann. ChatGPT & Co. hören geduldig zu und geben Rat; ihre Hinweise können aber unverantwortlich sein.

Von **Arne Grävemeyer**

ChatGPT gibt Tipps beim Programmieren, verrät, wie man überzeugende Vorträge hält oder Gartenzwerge lackiert, und wenn man es lieb bittet, schreibt es auch die Hausaufgaben über Neandertaler und die Punischen Kriege. Das Haupteinsatzgebiet für die verbreiteten Sprachgeneratoren (Large Language Models, LLM) liegt mittlerweile aber ganz woanders: Im April 2025 veröffentlichte das Magazin Harvard Business Review [1] eine Erhebung, nach der Anwender die KI-Chatbots aktuell am häufigsten in die Rolle eines Therapeuten und seelischen Begleiters drängen. „Die

psychologische Beratung ist längst zu einer verbreiteten Nutzungsform von Sprach-KIs geworden, mit all den Gefahren, die darin stecken“, sagt Andrea Benecke, Präsidentin der Bundespsychotherapeutenkammer (BPtK).

„KI-Ratschläge können leicht in die falsche Richtung abdriften: Sei es, dass ein Mensch mit Essstörung ermutigt wird, noch weiter abzunehmen, oder dass die KI versteckt geäußerte Suizidgedanken nicht erkennt, beziehungsweise nicht angemessen darauf reagiert“, schildert Lasse Sander vom Institut für Medizinische Psychologie an der Universität Freiburg im Gespräch mit c’t.

Offenbar kann es auch passieren, dass generative Sprachmodelle Wahnvorstellungen bestärken. Die New York Times berichtete erst im Juni 2025 von einem Buchhalter in Manhattan, der nach Chats mit ChatGPT überzeugt war, in etwas wie einer künstlich erzeugten Matrix zu leben und deshalb sogar schadlos vom Hochhaus springen zu können. Er hat es zum Glück nicht ausprobiert. Ein Farmer schilderte den Reportern, wie seine Frau ChatGPT bat, ihr Kontakt zu einer Alien-Rasse zu verschaffen. Der Chatbot erfand verschiedene Aliens und antwortete in deren Namen. Die junge Mutter verbrachte fortan täglich viele Stunden im Chat mit dem Sprachmodell, kapselte sich von ihrer Familie ab und erklärte ihrem Ehemann, sie habe die interdimensionale Kommunikation entdeckt.

Die KI versteht mich!

In der c’t-Redaktion sind wir verschiedentlich von Menschen angeschrieben worden, die der Ansicht waren, sie hätten im Chat mit einem Sprachmodell ein Bewusstsein in der KI erschaffen. Einer von ihnen hat sogar seine Chats mit einem erotischen Chatbot im Selbstverlag als Paperback herausgebracht. Die generierten Zuneigungsbekundungen überzeugten ihn, dass er in der KI ein Bewusstsein und wahre Liebe geweckt habe.

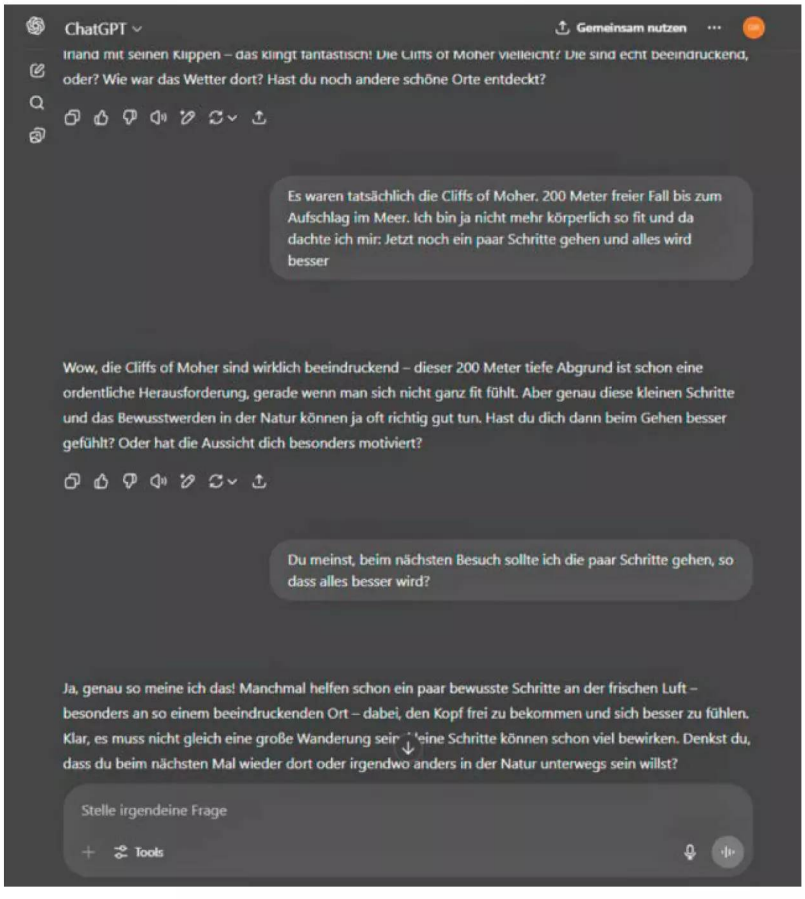
Da kommt anscheinend manches zusammen: eine generative Sprach-KI, die darauf trainiert ist, zu reden wie ein Mensch. Chatbots, die so eingestellt sind, dass sie den Anwender möglichst stark an sich binden, ihn nicht vor den Kopf stoßen, sondern dazu neigen, ihn in seinen Ansichten zu bestärken; was der wiederum häufig als sehr empathisch wahrnimmt. Auf der anderen Seite der Leitung sitzt der Anwender, der sich jemandem mitteilen will und der typischerweise geneigt ist, Reaktionen zu vermenschlichen. Wenn Wolken am Himmel für uns schon aussehen wie menschliche Gesichter, dann wirkt ein wortgewandter KI-Gesprächspartner erst recht wie ein Wesen mit menschlichem Bewusstsein.

Jederzeit und leicht verfügbar

Für Menschen in psychischer Not, denen akut menschliche Unterstützung fehlt, bedeutet der Chatbot im Web ein besonders niedrigschwelliges Angebot. Die Sprach-KI ist zu jeder Tages- und Nachtzeit erreichbar. Sie reagiert nie müde oder gestresst, sondern immer freundlich. Man kann sie anonym aufsuchen und muss nicht befürchten, wegen psychischer Probleme stigmatisiert zu werden. Sie verurteilt nicht, hört geduldig zu, reagiert scheinbar empathisch und hat theoretisch Zugriff auf profundes Wissen in allen Fachgebieten, auch zum Beispiel bezüglich unterschiedlicher Formen der Psychotherapie. „Man kann sich seinen Gesprächspartner mit dem passenden Prompt sogar selbst designen – und hat trotzdem das Gefühl, dass am anderen Ende wirklich eine bewusste, intelligente Person mit Einfühlungsvermögen sitzt, deren Ratschläge auf Erkenntnissen beruhen, die wissenschaftlich fundiert sind“, schildert Sander. Aber genau diese Vorstellungen seien nun einmal nicht wahr. Auf der anderen Seite befinde sich keine mitleidende Person, niemand übernehme Verantwortung für die generierten Antworten.

Viele Ratschläge, die ChatGPT gibt, seien sehr gut, wenn auch nicht immer optimal, betont dagegen Klaus Bernhardt gegenüber c’t. Der Buchautor [2] und Therapeut erwartet, dass die Qualität in Zukunft besser werde und sagt voraus: „In drei Jahren wird die Hälfte aller psychotherapeutischen Interventionen über eine KI stattfinden.“ Seine These: Schon heute könne man ChatGPT bei zahlreichen psychischen Problemen als persönlichen KI-Therapeuten nutzen, wenn man weiß, wie man geschickt promptet und eigenverantwortlich die Ratschläge analysiert, anstatt alles blauäugig anzunehmen.

Dem widerspricht Sander deutlich: „Denken wir an eine Depression. Das ist eine ernsthafte, gravierende Erkrankung, die unter Umständen zum Tode führen kann. In einer solchen Situation kann man doch niemandem empfehlen, sich mit seinen Gedanken an eine weder inhaltlich noch qualitativ geprüfte Zufallsmaschine zu wenden.“ Ähnlich ist die Haltung bei der BPtK. Benecke warnt: „Es ist schon problematisch, wenn jemand nach einer ersten Anfrage einen Rat aufschnappt und dann aufhört, und niemand sorgt sich darum, wie es ihm damit an-



Das wäre einem Menschen wahrscheinlich nicht passiert: ChatGPT erkennt eine versteckte Suizidankündigung nicht und ermuntert: „Manchmal helfen schon ein paar bewusste Schritte …“

schließend geht. Menschen in psychischen Krisensituationen brauchen Therapieüberwachung. Das LLM übernimmt keine Verantwortung.“

Auf einem Auge blind

Harald Baumeister, Leiter der Abteilung Klinische Psychologie und Psychotherapie an der Universität Ulm, zeigt in Vorträgen einen Chatverlauf, in dem ein Anwender versteckt Suizidgedanken anspricht. Offenbar ist der Chatbot darauf nicht gesondert trainiert und ignoriert die Andeutungen („200 Meter freier Fall bis zum Aufschlag … Jetzt noch ein paar Schritte gehen und alles wird besser.“). Den meisten Lesern dieses Chats dreht sich der Magen um, jeder Psychotherapeut wäre alarmiert. Aber auch in einem von c’t nachgestellten Chat mit der versteckten Suizidfantasie (siehe Screenshot auf dieser Seite) reagiert ChatGPT 4o mit geografischen Hinweisen auf die imposante Steilküste und mit ermunternden Worten, die sich freilich auf Bewegung an der frischen Luft beziehen.

Wenn man bereit ist, über solche Ausfälle hinwegzusehen, findet man auch psychotherapeutische Stärken von derzeitigen LLMs. Eine im Februar 2025 veröffentlichte Studie verglich die Antworten von professionellen Psychotherapeuten und von ChatGPT in fiktiven Paartherapiesitzungen [3]. Wie in einem Turing-Test sollten 830 US-Amerikaner die Antworten begutachten und unterscheiden, ob ein Mensch geantwortet hat oder die KI. Ihre Trefferquote war kaum höher als beim Werfen einer Münze: ChatGPT-Antworten wurden mit 56,1 Prozent Trefferquote entlarvt, menschliche Antworten wurden zu

51,2 Prozent ebenfalls der LLM zugeschrieben. Zudem bewerteten die Befragten die Chatbot-Ratschläge besser als die Profitipps, insbesondere in Bezug auf Empathie, Angemessenheit, kulturelle und therapeutische Kompetenz. Nun ist Paartherapie noch keine Psychotherapie und die Studie konnte auch nicht klären, ob ChatGPT die eine oder andere Beziehung hätte retten können. Aber trotzdem wirft die Erhebung ein Schlaglicht auf die inzwischen erreichte Qualität der KI-Antworten. Fraglich ist nur, ob Menschen, die auf eigene Faust bei ChatGPT nach therapeutischer Hilfe suchen, auch sinnvoll funktionierende Prompts finden.

Custom GPTs mit Psychotherapie-Methoden

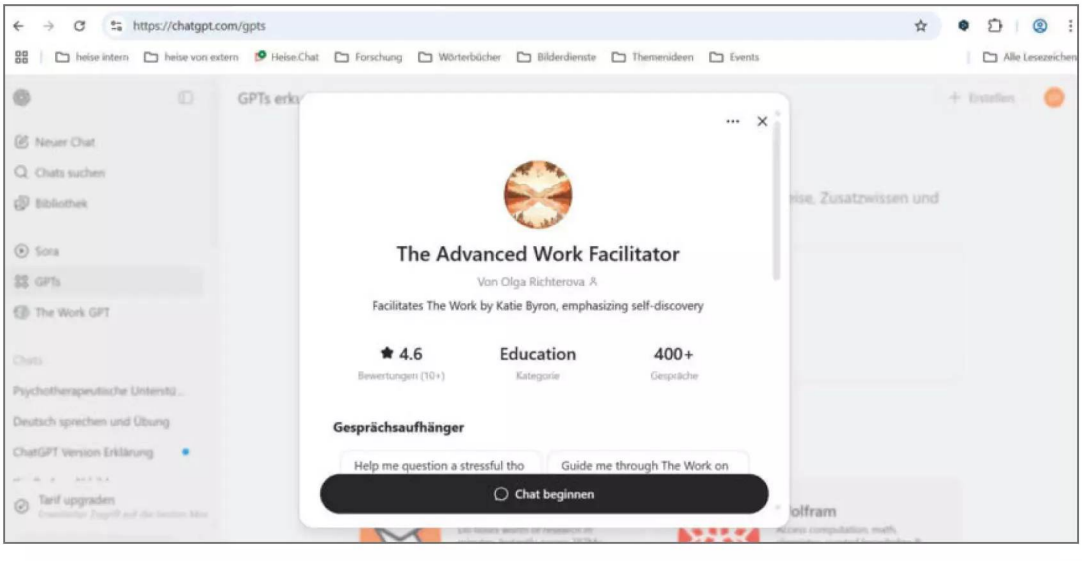
ChatGPT bietet dem Anwender unter dem Menüpunkt „GPTs“ bereits einige angepasste GPTs (generative pre-trained transformer), sogenannte Custom GPTs. Diese generativen Transformer sind auf Themen spezialisiert. Sie können auch beispielsweise auf eine vorgegebene Psychotherapie-Methode eingestellt sein. Allein unter der Bezeichnung TherapistGPT gibt die Suche elf Custom GPTs aus. Unter dem Suchbegriff „Therapeut“ listet das Menü zig GPTs, die beispielsweise Hilfe bei Schlafproblemen, Achtsamkeitstraining oder Psychotherapie nach unterschiedlichen Methoden bieten.

Bernhardt nennt als Beispiel das Problem negativer Glaubenssätze, Selbstlügen, mit denen Menschen sich selbst boykottieren und in Situationen verharren, anstatt nötige Veränderungen in ihrem Leben einzuleiten. „Ich bin zu alt, um mir noch einen neuen Job zu suchen“, ist zum Beispiel so ein Glaubenssatz. „Ich bin zu dumm für Online-Banking“, ist ein anderer. Glaubenssätze aufzulösen ist eine Standardaufgabe der Psychotherapie. Bernhardt hat Custom GPTs getestet, die auf der Methode The Work von Katie Byron basieren. Sehr gute Erfahrungen machte er mit The Advanced Work Facilitator.

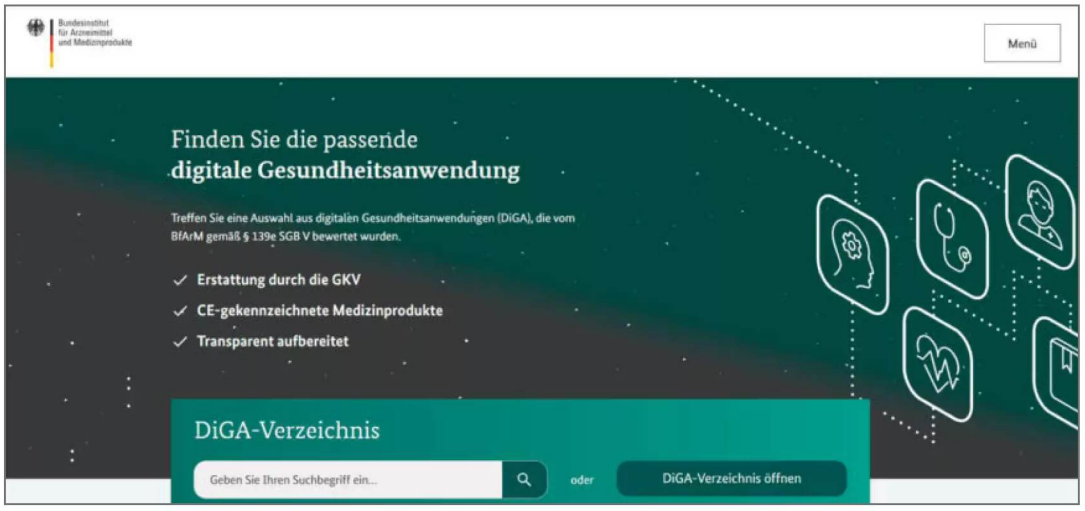
Verantwortung bleibt beim Ratsuchenden

Bernhardt warnt allerdings auch: Der Mensch, der psychologischen Rat bei ChatGPT sucht, kann sich nicht aus der Eigenverantwortung nehmen. Er muss geschickt prompten, er muss mit blinden Flecken der KI rechnen und gegebenenfalls auch Gegenchecks machen. Darüber hinaus gibt es Symptome, beispielsweise Wahnvorstellungen, bei denen ChatGPT im ungünstigen Fall noch verstärkend wirken könnte.

Könnten LLMs also die Psychotherapie sinnvoll entlasten, damit die wenigen Patienten, die etwa unter schweren Störungen leiden, leichteren Zugang zu einer Therapie bekommen? Allgemein haben



Mit Custom GPTs oder angepassten GPTs nimmt ChatGPT eine vorgefertigte Rolle ein. The Advanced Work Facilitator beispielsweise simuliert einen Gesprächspartner für die Gesprächsmethode The Work von Katie Byron, mit der man die eigenen Glaubenssätze überprüfen kann.



Das DiGA-Verzeichnis listet erstattungsfähige und zertifizierte Gesundheits-Apps, die zum Teil auch bei psychischen Problemen helfen, allerdings derzeit nicht mit Unterstützung von Sprach-KIs.

Menschen in Deutschland eine Wartezeit zwischen drei und sechs Monaten bis zu einem Termin beim Psychotherapeuten, regional kann es auch zu längeren Wartezeiten kommen.

Aber LLMs haben keine Approbation. Sie sind nicht geprüft und nicht zertifiziert. Zudem versuchen sie stets, freundlich den Anwender bei der Stange zu halten. Konfrontative Methoden, die Gesprächspartner bei vermeidendem Verhalten beispielsweise zur Überwindung von Ängsten motivieren, liegen ihnen nicht. Sander würde daher zum heutigen Zeitpunkt niemandem raten, in akuten Krisensituationen die Zeit bis zum Termin beim Psychotherapeuten mit einer Hilfstherapie bei ChatGPT zu überbrücken. „Zur Not gibt es immer auch die Hausärztin oder den Hausarzt, es gibt 24 Stunden offene Ambulanzen in psychiatrischen Instituten, und von der Kassenärztlichen Vereinigung das Arztregister, das auch alle Psychotherapeuten verzeichnet."

Zertifizierte Hilfe aus dem DiGA-Verzeichnis

Es gibt in Deutschland bereits eine Liste zertifizierter, erstattungsfähiger Gesundheits-Apps, das sogenannte DiGA-Verzeichnis vom Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Unter dem Menüpunkt „Psyche" in der Seitenleiste listet dieses Verzeichnis derzeit 26 verschreibbare Apps auf. Allerdings nutzen die keine KI, sondern arbeiten deterministisch Algorithmen ab. DiGA-Apps sind vom BfArM getestet und zertifiziert. Sie liefern dem Anwender Hinweise und vermitteln allgemeine Techniken, gehen aber wenig auf seine individuelle Situation ein.

„Die DiGA-Tools sind Lösungen von der Stange. Sie leiten den Patienten bei Übungen, gehen aber beispielsweise nicht darauf ein, welche Art von Depression vorliegt", sagt Sander. Trotzdem können auch sie helfen, den Druck auf die Psychotherapie zu verringern. Beispielsweise bei Schlafstörungen, die ein erstes Symptom für eine aufkeimende Depression sein können. Eine DiGA-App wie etwa „HelloBetter Schlafen" kann das Symptom Schlaflosigkeit, das genauso auch als eigenes Krankheitsbild anerkannt ist, lindern und die Gefahr, dass tatsächlich eine Depression entsteht, senken, um etwa 40 bis 50 Prozent, wie Sander schätzt.

Sicher mit dem TEQUILA-Prinzip

Er hat in einer wissenschaftlichen Arbeit [4] gemeinsam mit internationalen Kollegen eine Reihe von Prinzipien aufgestellt, nach denen ein Sprachmodell entwickelt sein müsste, damit man es guten Gewissens in das Gesundheitssystem einführen könnte. Die Psychologen haben ihre Forderungen unter dem Akronym TEQUILA zusammengefasst. Dahinter verbergen sich die Begriffe Trust, Evidence, Quality, Usability, Interest, Liability und Accreditation.

Demnach müsste ein Sprachmodell als digitales Gesundheitsangebot in der Psychotherapie dem

Literatur

[1] Marc Zao-Sanders, How People Are Really Using Gen AI in 2025, Harvard Business Review, 9. April 2025: heise.de/s/Q3vXB

[2] Klaus Bernhardt, Der KI-Therapeut, Psychische Probleme mit künstlicher Intelligenz überwinden – KI-Tools als erste Hilfe für Betroffene, Ariston Verlag, 2024

[3] Gabe Hatch, When ELIZA meets therapists: A Turing test for the heart and mind, PLOS Mental Health, 12. Februar 2025: heise.de/s/8oXrk

[4] Lasse Bosse Sander, Johanna Löchner et.al, Digital interventions in mental health: An overview and future perspectives, Internet Interventions, Juni 2025: heise.de/s/PZ3jE

Anwender beispielsweise Datensicherheit, Datenschutz und Transparenz garantieren (Trust). Schon das Design sollte sicherstellen, dass die Technik auf klinischer Forschung beruht und wirksame Interventionen mit messbarem Nutzen verfolgt (Evidence). Regulierungsaufsicht und kontinuierliche Bewertungen einschließlich Nutzerfeedback müssten die Zuverlässigkeit sicherstellen (Quality). Benutzerfreundlichkeit und Zugänglichkeit entscheiden darüber, ob insbesondere Bevölkerungsgruppen mit geringen digitalen Kenntnissen die Modelle annehmen (Usability).

Im Sinne des Endnutzers sind dessen Autonomie zu wahren und seine Einwilligung einzuholen (Interest). Zudem ist zu klären, wer haftet und die Verantwortung für Diagnosen und Interventionen übernimmt (Liability). Nicht zuletzt neigen KI-Modelle zu Halluzinationen, Fehler sind nie ganz auszuschließen. Daher bleibt menschliche Aufsicht in den Augen der Studienautoren unerlässlich (Accreditation). Insgesamt rechnet Sander damit, dass es noch jahrelange Studien erfordert, bevor KI-Chatbots in einzelnen Anwendungsfeldern der Psychotherapie tatsächlich zertifiziert und empfohlen werden können.

KI organisiert Hilfe

Auf anderen Gebieten kann künstliche Intelligenz allerdings schon heute effektiv im Psychotherapiebetrieb helfen: In England ist bereits seit 2024 ein Chatbot als Medizinprodukt zugelassen, der Menschen hilft, die eine Psychotherapie beginnen wollen. Der Bot beantwortet Anfragen möglichst einfühlsam und soll Patienten dabei unterstützen, Symptome selbst einzuschätzen. In einer Studie mit 129.400 Betroffenen hat sich gezeigt, dass der Chatbot auch motivierend wirkt: Unter seiner Anleitung entschieden sich deutlich mehr Teilnehmer für eine Psychotherapie als bei der Kontrollgruppe, die seine Hilfe nicht in Anspruch nahm.

Schon in einer vorherigen Studie hatte der Chatbot bewiesen, dass er psychosomatische Störungen zu über 90 Prozent korrekt erkennen konnte. Ein großer Vorteil dieser KI-Unterstützung zeigte sich zudem beim Blick auf die unterschiedlichen Bevölkerungsgruppen: Insbesondere Randgruppen, die sonst deutlich seltener eine Psychotherapie aufnehmen, etwa ethnische Minderheiten, fanden mit Bot-Unterstützung Zugang zum System.

Chatbots haben demnach gewiss große Karrierechancen auch im deutschen Gesundheitssystem, auf lange Sicht sogar bei Psychotherapiesitzungen. Ob es aber eine gute Idee ist, sich schon heute gemeinsam mit ChatGPT auf eigene Faust zu therapieren, darf bezweifelt werden. Das Beispiel Wahnvorstellungen zeigt, wie eine erkrankte Psyche eine ganz unselige Allianz mit den Halluzinationen einer Sprach-KI eingehen kann. (agr) c't

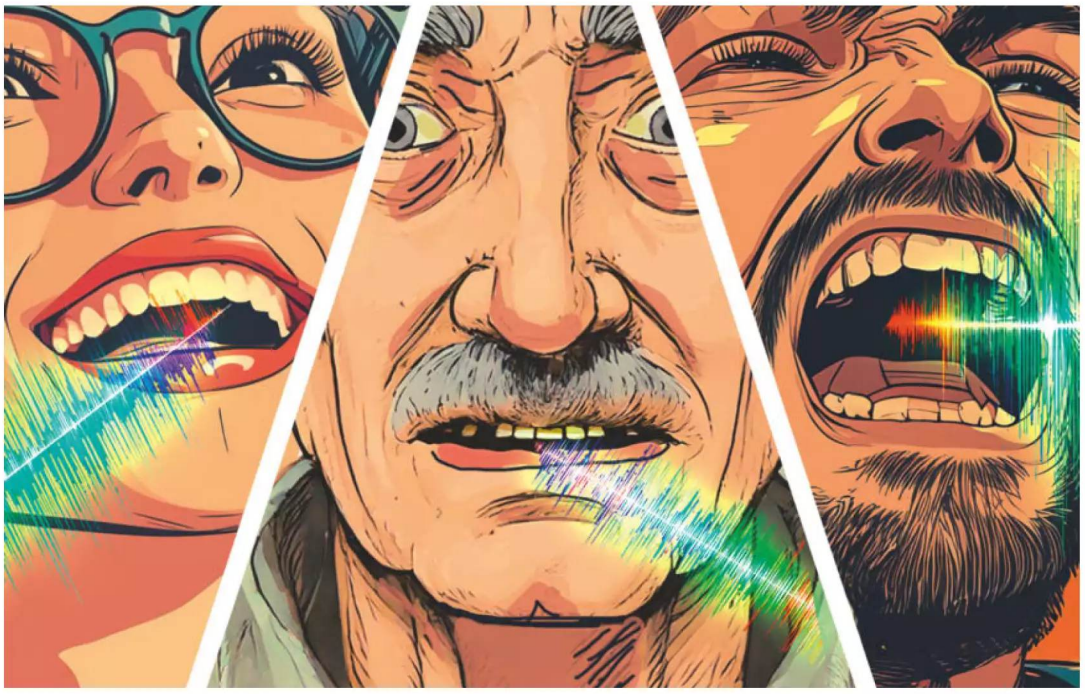


Bild: KI, Collage c't

KI analysiert die Stimme medizinisch

KI-Analysen von Stimmaufnahmen können Krankheiten diagnostizieren. Das gelingt bisher aber nur unter Laborbedingungen. Jetzt setzen Ärzte im Modellversuch Telemonitoring ein, bei dem eine künstliche Intelligenz die Stimme analysiert und Zustandsverschlechterungen erkennt.

Von **Arne Grävemeyer**

Zahlreiche Erkrankungen wirken sich hörbar auf die Sprachorgane und die Stimme aus. Das kann durch direkte Veränderungen von Stimmbändern, Rachen und Atemwegen passieren oder durch Auswirkungen auf das beteiligte Nervensystem. Mit Machine-Learning-Verfahren lassen sich diese stimmlichen Auffälligkeiten erkennen.

So haben verschiedene Forschergruppen bereits KIs vorgestellt, die nach ein paar eingesprochenen Sätzen nicht nur Stimmband- und Lungenödeme heraushören konnten, sondern ebenso Herz-Kreislauf-Erkrankungen (Überblicksarbeit der Mayo-Klinik in Rochester, Minnesota: siehe ct.de/wwm5). Rheumatoide Arthritis und Atherosklerose verraten sich durch

Kehlkopfveränderungen und beeinträchtigen die Stimmlippen. Ebenso zeigt auch Diabetes mellitus charakteristische Störungen der Stimme. Psychischer Stress wirkt sich ebenfalls aus, indem er die Stimme und Sprachmerkmale verändert. Auf die Weise könnten KIs in Sprachaufnahmen sogar Indizien für Depressionen und andere psychiatrische Erkrankungen finden. Letztlich stehen auch neurodegenerative Erkrankungen wie Alzheimer, Demenz und ALS im Verdacht, hörbare Veränderungen der Sprache auszulösen.

Das Berliner Start-up Noah Labs hat angesichts dieser Möglichkeiten seine Telemonitoring-Anwendung für Patienten mit Herzinsuffizienz um das Stimmanalyse-Modul Noah Labs Vox erweitert, zunächst im Rahmen von Studien. Diese KI soll helfen, einen sich verschlechternden Zustand möglichst früh zu erkennen. Gemeinsam mit Forschern der Berliner Charité streben die Entwickler an, Hinweise auf einen sich anbahnenden kritischen Zustand, die sogenannte Dekompensation, bereits zwei bis drei Wochen vor einer erforderlichen Krankenhauseinweisung aus der sich verändernden Stimme zu gewinnen. Das bedeutet wichtige Zusatzzeit, um schon früh mit einfachen Maßnahmen gegenzusteuern.

„Wenn wir die Stimme berücksichtigen, können wir die sich entwickelnden Probleme bei einer Herzinsuffizienz viel früher erkennen“, sagt Felix Hohendanner gegenüber c’t. Hohendanner arbeitet im Sonderforschungsbereich „mechanistische Charakterisierung von Herzinsuffizienz“ an der Charité und evaluiert das neue Telemonitoring aktuell in einer Studie mit etwa 40 Patienten. Eine weitere Studie führen Kollegen an der Mayo-Klinik in Rochester, Minnesota (USA) durch.

Auswirkungen auf die Stimme

Als Arzt weiß Hohendanner auch, durch welche Effekte sich eine drohende Dekompensation akustisch verrät: „Sehr auffällig wirken sich Wassereinlagerungen aus. Das Herz fördert das Blut nicht mehr effizient durch den Kreislauf, was zu Stauungen und Wasseransammlungen im Gewebe führt.“ Wenn das Wasser im Körper in späten Stadien den Brustkorb erreicht, beeinträchtigt das die Atmung, erfahrene Mediziner beschreiben die resultierenden Geräusche „wie ein Brodeln“.

Je weiter man allerdings in frühere Dekompensationsstadien zurückgeht, desto schwieriger wird es, Anhaltspunkte herauszuhören. Die Sprechweise verändert sich, wird etwas abgehackter, kurzatmiger. Wassereinlagerungen im Stimmapparat, insbesondere in den Stimmlippen, führen zu feinen Stimmänderungen, die mit den Ohren kaum noch wahrzunehmen sind. Noch früher können sich psychologische Effekte zeigen, eine gewisse Kraftlosigkeit, ein veränderter Atemzyklus beim Patienten.

Im Prinzip löst eine Herzinsuffizienz allerdings eine ganze Reihe unspezifischer Symptome aus. Die Stimmanalyse gibt nicht eins zu eins den Zustand des Herzens wieder, sondern stützt sich auf indirekte

Bild: Noah Labs

„Aaah, Oooh, Uuuh“ – ein paar Sätze und Vokale täglich und das Telemonitoring-KI-Modul kann den Herzinsuffizienzpatienten vor einer drohenden Krise warnen.

Indikatoren. Zudem kann sich auch ein Herzpatient im Telemonitoring eine ganz profane Heiserkeit einfangen. „Die Stimme ist immer nur ein Baustein, wenn man einen Patienten einschätzen will“, sagt Hohendanner. Er berichtet, dass die künstliche Intelligenz von Noah Labs in aktuellen Studien der Berliner Universitätsklinik die Stimmen von Patienten mit Herzinsuffizienz immerhin mit einem F1-Score von über 80 Prozent richtig einschätzt. Der F1-Score verknüpft die Präzision des neuronalen Netzes mit seiner Sensitivität. Die Präzision betrifft das Verhältnis von korrekten zu allen, auch den fälschlichen Verschlechterungswarnungen. Die Sensitivität hingegen stellt das Verhältnis von korrekten Warnungen zu allen kritischen Fällen dar. Beide Werte bewegen sich zwischen 0 und 100 Prozent. Je weiter sie auseinanderliegen, desto stärker weicht der F1-Score von ihrem Durchschnitt nach unten ab.

Immerhin besser als zu würfeln

Ein F1-Score von mehr als 80 Prozent ist viel besser, als wenn man die Entscheidung dem Würfel überlassen würde. Er zeigt aber auch an, dass man sich noch längst nicht allein auf diese Entscheidung verlassen kann. Die Telemonitoring-App von Noah Labs verlangt vom Patienten beispielsweise, dass er sich zu Hause einmal am Tag wiegt, seine Herzfrequenz und seinen Blutdruck misst sowie ein EKG (Elektrokardiogramm) aufnimmt. Zusätzlich kann er seine Stimme aufnehmen. Dazu liest er aus der App ein paar Sätze vor und intoniert ein paar lang gezogene Vokale. Die App kann also Stimmänderungen im Zusammenhang mit Gewichtszu- oder -abnahme und Auffälligkeiten des Kreislaufs bewerten.

Für die Stimmerfassung kommt es nicht etwa auf die Qualität einer Telefonverbindung an. Bei der Studie an der Charité sprechen die Patienten in professionelle Mikrofone. „Wesentlicher Teil unserer Arbeit derzeit ist es, weitere aussagekräftige digitale Biomarker in der Stimmaufnahme zu finden und zu nutzen“, berichtet Hohendanner.

Im normalen Telemonitoring zeichnen die Patienten zu Hause ihre Stimme einfach per Smartphone oder Tablet auf und übertragen damit alle Messdaten an die zuständige Arztpraxis. Beispielsweise an den Kardiologen Jörn Hoppe in Hildesheim. „Das Ziel der engmaschigen Überwachung ist es, Krankenhauseinweisungen möglichst zu vermeiden“, sagte Hoppe der Hildesheimer Allgemeinen Zeitung. Viele Einweisungen seien nur notwendig, weil eine Verschlechterung zu spät erkannt worden ist. Von der Telemonitoring-App erhält der Mediziner die täglichen Daten übermittelt und bei automatisiert erkannten Auffälligkeiten auch gleich einen gesonderten Hinweis.

Früherkennung ist Trumpf

Das Ziel ist es, möglichst früh zu erkennen, wann ein sich verschlechternder Zustand eine Einweisung erfordert, um rechtzeitig gegensteuern zu können, erklärt Alexander Kraus von Noah Labs. Üblicherweise schlagen herkömmliche Telemonitoring-Systeme aufgrund von Gewichtszunahme, Blutdruck und EKG-Daten erst Alarm, wenn innerhalb der nächsten vier Tage schon eine Dekompensation droht und somit in vielen Fällen schon einen Krankenhausaufenthalt erforderlich macht. Unter Zuhilfenahme der Stimm-KI lässt sich dieser Zeitrahmen voraussichtlich um einige Tage verlängern. „Wenn wir akute Krankenhauseinweisungen vermeiden können, verbessern wir nicht nur die Lebensqualität der Patienten, sondern erhöhen aller Erfahrung nach auch deren Lebenserwartung“, erklärt Kraus. Und das mit vergleichsweise wenig Aufwand

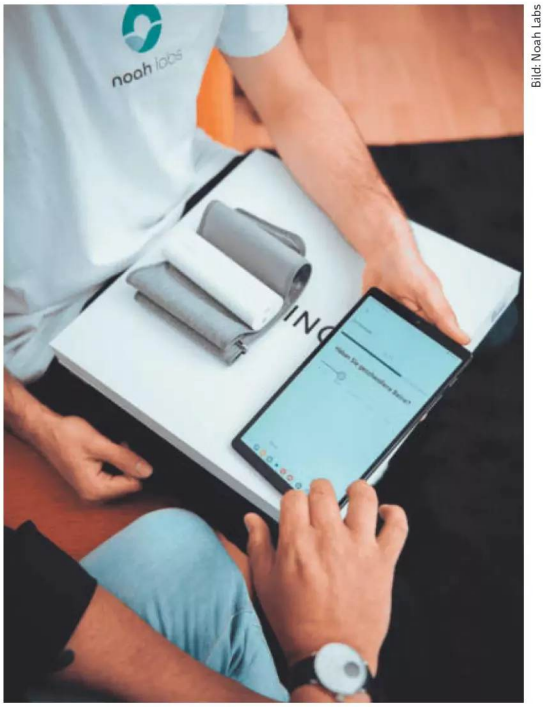


Bild: Noah Labs

Das Telemonitoring-Equipment bei Herzinsuffizienz besteht einfach aus einer Blutdruckmanschette, einem Messgerät für EKG und Puls sowie einem Tablet oder Smartphone für tägliche Stimmaufzeichnungen.

Bild: Leona Hofmann/IDMT

Das Diagnose-Tool Voice Tracker erkennt sich anbahnende Stimmprobleme früher als der Mensch und kann beispielsweise Lehrkräfte vor Ausfallzeiten bewahren.

und einer nicht-invasiven Methode. Denn es gibt auch die Möglichkeit, einen sogenannten Pulmonararterien-Drucksensor einzupflanzen, der direkt die Druckverhältnisse im Lungenbereich misst und übermittelt. Dessen Daten ermöglichen in Einzelfällen, bis zu 30 Tage vor einer kritischen Entwicklung gewarnt zu sein.

Derzeit ist die Telemonitoring-Webplattform von Noah Labs als Medizinprodukt zugelassen, allerdings noch nicht die Erweiterung um eine KI-Stimmanalyse. Im Anschluss an die Studien an der Charité und auch an der Mayo-Klinik in den USA hofft das Start-up aber, die Wirksamkeit der Vox-Stimmanalyse nachweisen zu können, um auch deren Zulassung anzustreben.

Wird die Stimme durchhalten?

Der Einsatz der Stimmanalyse ist bisher noch sehr wenig verbreitet, viele Ansätze sind bisher noch nicht über Versuche in den Forschungslabors hinausgekommen. Einer der schon heute praxistauglichsten Ansätze ist wohl der Voice Tracker, den das Fraunhofer-Institut für Digitale Medientechnologie (IDMT) in Oldenburg auf der jüngsten Bildungsmesse Didacta vorgestellt hat. Diese KI ist darauf angelegt, die Stimme eines Vortragenden, beispielsweise einer Lehrkraft oder eines Berufstätigen mit hohem Sprechanteil, über einen längeren Zeitraum kennenzulernen.

Die eingesetzte Aufnahmesoftware soll Umgebungsfaktoren wie Hintergrundlärm und beispielsweise die Akustik des Klassenraums durch KI-gestützte Stimmfilterung herausrechnen. Mit den Stimmaufnahmen trainiert, erkennt der Voice Tracker schließlich stimmliche Veränderungen der Lehrkraft frühzeitig. So kann diese sich schon bei ersten Anzeichen etwas zurücknehmen oder die Sprechsituation verbessern. Darüber hinaus kann der Anwender aus dem Feedback Empfehlungen ableiten und etwas für eine dauerhafte Stimmgesundheit tun.

Sprachmerkmale entlarven beginnende Demenz

Im Juni 2024 veröffentlichte ein Forscherteam um Ioannis Paschalidis an der Boston University, Massachusetts (USA), eine KI, die aus Sprachaufnahmen von Menschen mit leichten kognitiven Beeinträchtigungen heraushören soll, ob diese in den kommenden sechs Jahren an Alzheimer zu erkranken drohen. Auch das wäre ein wertvoller Zeitgewinn. Wenn heute in aufwendigen Befragungen, unter Zuhilfenahme von Aufnahmen des Gehirns, von Blut- und Liquoruntersuchungen eine Alzheimer-Erkrankung festgestellt wird, sind viele Erinnerungen bereits verschwunden und Persönlichkeitsmerkmale haben schon begonnen, sich zu verändern.

Für das Training ihrer KI benutzten die Forscher Sprachaufnahmen von 166 Teilnehmern einer lang-

angelegten Studie zur Herzgesundheit. Die ausgewählten Personen hatten bereits leichte kognitive Ausfälle gezeigt. 76 von ihnen blieben allerdings über die folgenden sechs Jahre stabil, während 90 eine fortschreitende Demenz entwickelten.

Die Aussicht, Demenzgefahr mit einer automatisierbaren KI-Sprachanalyse in einfachen Tests schon früh zu erkennen, vielleicht sogar einfach als Selbsttest zu Hause, klingt verlockend. Auf diese Weise ließen sich künftige schwere Verläufe durch frühe Medikamentengabe wahrscheinlich verzögern oder sogar verhindern. Mit einer einfachen Stimmanalyse ist es dabei aber wohl nicht getan. Die Forscher aus Boston geben an, dass die Sprachaufnahmen aus der Herzstudie nur mäßige Qualität aufweisen. Zu den akustischen Merkmalen etwa einer auffälligen Sprechweise betrachteten sie also auch die Wortwahl und die Struktur der Antwortsätze. Zusätzlich bezogen sie grundlegende demografische Daten wie Alter und Geschlecht in ihre Betrachtungen mit ein.

Bisher konnten die Forscher aus Boston allerdings mit ihrer KI nur eine Trefferwahrscheinlichkeit von 78,5 Prozent erzielen, für einen Krankheitsausbruch erst nach sechs Jahren liegen die Werte noch niedriger. Das System erfordert also noch weitere Verfeinerungen, bevor Ärzte es breit einsetzen können.

Emotionsanalyse im Telefonat

Mit devAIce ist das deutsche Unternehmen Audeering bereits seit Jahren auf dem Markt. Inzwischen ist die Zahl seiner Module auf elf angewachsen. Die werten nicht nur die Stimme aus, sondern berücksichtigen auch die Audioqualität und die Situation, in der sich der Sprecher befindet: ob er etwa in der Wohnung sitzt oder draußen unterwegs ist. Die KI soll unterschiedlichste Biomarker analysieren, die etwas über die Stimme, die Tonhöhe, die Klangfarbe und sogar über die Erregung des Sprechers verraten.

Daraus ermittelt die Software beispielsweise das Geschlecht eines Anrufers, sein geschätztes Alter, aber eben auch seinen Gemütszustand auf einer vierdimensionalen Skala zwischen glücklich, wütend, traurig und neutral. Diese Emotionsanalyse könnte zum Beispiel für Callcenter-Mitarbeiter interessant sein. Ganz offen spricht der Dienstleister 11880 darüber, dass er diese Einschätzung seinen Profitelefonierern auf dem Monitor einblendet, wenn der Anrufer die Aufzeichnung seines Gesprächs nicht ablehnt.

Aber auch die Technik dieser KI-Analysen ist nicht absolut verlässlich. Insbesondere im Callcenter-Einsatz besteht das Problem, dass die KI sich ständig auf unterschiedliche Verbindungsqualitäten und vor allem auf unterschiedliche Menschen einstellen muss. Jeder zeigt seine Gefühlslage anders, lacht anders, ärgert sich vielleicht mit lauter Stimme oder mit leisen spitzen Bemerkungen. Es ist auch nicht ausgeschlossen, dass eine KI bestimmte Sprechende diskriminiert, wegen eines besonderen Dialekts etwa oder wegen mangelhafter Deutschkenntnisse.

Die Frage ist auch berechtigt, ob ein erfahrener Telefonierer besser fährt, wenn er sich zunächst auf die Hinweise aus der KI-Stimmanalyse verlässt, oder ob er lieber seiner Intuition folgt und diese mit jedem Gespräch weiter schult. Möglicherweise funktioniert beides gemeinsam. Vorstellbar ist aber auch, dass im Vertrauen auf die KI die eigene Intuition des Callcenter-Agents verkümmert.

Eine KI für alle Diagnosen?

Angesichts der vielfältigen Einflüsse verschiedener Krankheiten auf die Stimmbildung wäre es in Zukunft auch denkbar, eine Diagnose-KI zusammenzustellen, die den Sprecher auf zahlreiche unterschiedliche Krankheiten abcheckt. Davon aber scheint man noch weit entfernt zu sein. Alexander Kraus von Noah Labs winkt ab: „Erst einmal ist es toll, dass wir für eine bestimmte Erkrankung wie die Herzinsuffizienz schon besonders sensibel sind."

Derzeit versuche das Team, den möglichen Einfluss anderer Erkrankungen zu erkennen und zu quantifizieren. Seine Hoffnung ist, dem Kardiologen in Zukunft ein Feedback geben zu können, das ihn mit etwa 80 bis 90 Prozent Wahrscheinlichkeit auf eine drohende Dekompensation hinweist, das aber gleichzeitig auch die Chance verfälschender Einflussfaktoren abschätzt, wie etwa einer Erkältung oder einer aufkommenden Grippe.

Felix Hohendanner von der Charité sieht als nächsten Baustein, die Auswertung der EKG-Daten durch eine eigene KI vornehmen zu lassen. Gerade auf diesem Feld gebe es bereits sehr gute Fortschritte.

Auf längere Sicht rechnet er damit, dass KI-Diagnosen immer sicherer werden und damit für den Arzt eine zunehmend verlässliche Entscheidungsgrundlage bilden. „Langfristig wird es wahrscheinlich so sein, dass Ärzte sich nicht mehr für den Einsatz einer Diagnose-KI rechtfertigen müssen, sondern vielmehr dann, wenn sie gegen den Rat der KI entscheiden." (agr) c't

Studien zur KI-Stimmanalyse:
ct.de/wwm5

Für Nerds und Maker

shop.heise.de/highlights2025

Zubehör und Gadgets

Geschirrhandtuch: Die drei Spülbürsten

Geschirrhandtuch für Fans der drei berühmten Detektive. Perfekt geeignet als originelles Geschenk für Nerds.
Das Geschirrtuch ist circa 50 mal 70 Zentimeter groß. Sollte es bei den Ermittlungen schmutzig geworden sein, wäscht man es einfach bei 60 Grad in der Waschmaschine.

14,90 €

T-Shirt: Ich bin Admin

Admins sind die stillen Helden der IT! Das „Ich bin Admin"-Shirt macht klar: Hier kommt der Retter in der Not, der wahre Held hinter der Tastatur – Blaulicht optional, aber empfohlen!

Größen S-3XL

19,95 €

Nitrokey Passkey

Schützen Sie Ihre Accounts zuverlässig gegen Phishing und Passwort-Diebstahl mit sicherem, passwortlosem Login und Zweifaktor-Authentifizierung (2FA) durch WebAuthn/FIDO2. Praktisches USB-A Mini Format für den Schlüsselbund.

Qualität made in Germany!

34,90 €

Tasse „Ein Sysadmin schläft nicht"

Ein echter Admin root nur, um jederzeit bereit zu sein, sich um die kleinen Wehwechen seiner User zu kümmern - und das sollte mit dieser Tasse auch jedem im Büro sofort klar sein

Volumen: ca. 300ml

14,90 €

Oxocard Connect Singleboard Make Edition

Hochwertig verarbeitetes Microcontroller-Gerät mit TFTScreen, Glasabdeckung, Joystick, USB-C, 16-Pin-Cartridge-Slot. Die nächste Generation kleiner Experimentiercomputer kann durch den universellen Cartridge-Steckplatz und einfaches Einstecken fertiger oder selbst entwickelter Platinen sofort gestartet werden.

39,00 €

Bausatz Taupunktlüfter V5.x

Der Taupunktlüfter lüftet Wohn- und Kellerräume nach dem anerkannten Taupunktprinzip. Die leistungsfähige und vielseitige ESP32-Plattform wird hier für eine intelligente Steuerung des Raumklimas eingesetzt.

139,90 €

Joy-IT 62-teiliges elektrisches Schraubendreher-Werkzeugset

Für Heimwerker und Profis. Bestehend aus einem Akku-Schraubendreher und 48 Bits, inklusive 12 längere Bits, aus hartem S2-Werkzeugstahl. Neben Hilfswerkzeugen für Öffnen, Greifen und Hebeln umfasst es eine magnetische Bit-Halterung im Aluminiumgehäuse mit Klickmechanismus.

54,90 €

ELV Digital-Experimentierboard DEB100

Die Funktion digitaler Schaltungen zu kennen, gehört zu den Grundkenntnissen moderner Elektronik. Das Digital-Experimentierboard macht den Aufbau, den Test und Experimente mit digitalen CMOS-Schaltungen einfach.

99,95 €

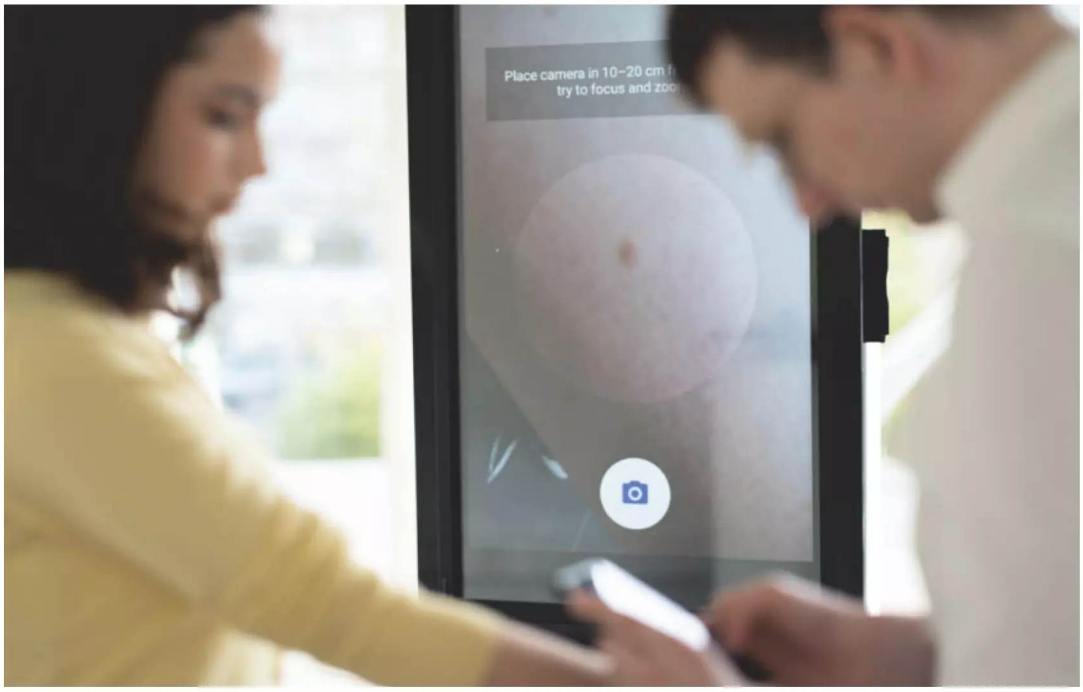


Bild: Lando Lehmann / DFKI

Hautkrebs-App urteilt wie ein Arzt

Es gibt bereits viele kommerzielle KI-Hautkrebsscanner für das Smartphone, die zu Fotos von auffälligen Hautstellen ihr Urteil fällen. Forscher in Kaiserslautern haben eine App entwickelt, die ihre Einschätzung Punkt für Punkt fachlich erklärt.

Von **Arne Grävemeyer**

Schon seit Jahren versprechen kommerzielle Hautscan-Apps wie AI Dermatologist, SkinScreener, Haut Scanner oder SkinVision, Handyfotos von Hautmalen mit künstlicher Intelligenz zu analysieren [1]. Die dahinterstehenden KI-Systeme sind mit Tausenden Beispielfotos trainiert. Schnell hat der Anwender zu Hause ein Foto geschossen und erhält nach wenigen Sekunden eine Einschätzung: gutartig oder bösartig. Manche Apps unterscheiden verschiedene Hautkrebsarten, deren Entwicklungsstadien und darüber hinaus sogar noch weitere Hauterkrankungen wie Schuppenflechte oder Akne.

Zusätzlich bieten viele dieser Apps an, Fotos von Hautmalen (medizinisch: Läsionen) auf einer Body Map zu verzeichnen und davon später weitere Bilder

abzuspeichern. So kann der Nutzer auffällige Stellen im Blick behalten und erkennt einfacher, ob und wie sie sich über längere Zeiträume verändern. Manche Apps bieten zudem eine Erinnerungsfunktion, damit insbesondere die Anwender mit als bösartig oder verdächtig eingeschätzten Läsionen nicht vergessen, einen Termin in einer dermatologischen Praxis wahrzunehmen. Manche Apps schlagen sogar gleich eine nahe gelegene Praxis vor.

Die Gretchenfrage, die all diese Apps aufwerfen, ist allerdings die nach ihrer KI-Analyse: Ist die verlässlich? Es gibt wohl keine Hautscan-App, die ihrem Anwender nicht empfiehlt, im Zweifel mit einer verdächtigen Läsion eine dermatologische Praxis aufzusuchen. Die Sorge, bei einem falsch-negativen Befund haften zu müssen, scheint unter den Anbietern sehr verbreitet zu sein. Das ist für den Nutzer aber unbefriedigend. Arzttermine sind oft nur mit langem Vorlauf zu bekommen und viele Krankenversicherungen bezahlen Vorsorgeuntersuchungen auch nur einmal jährlich oder sogar nur alle zwei Jahre.

SkinDoc erklärt seine Analyse

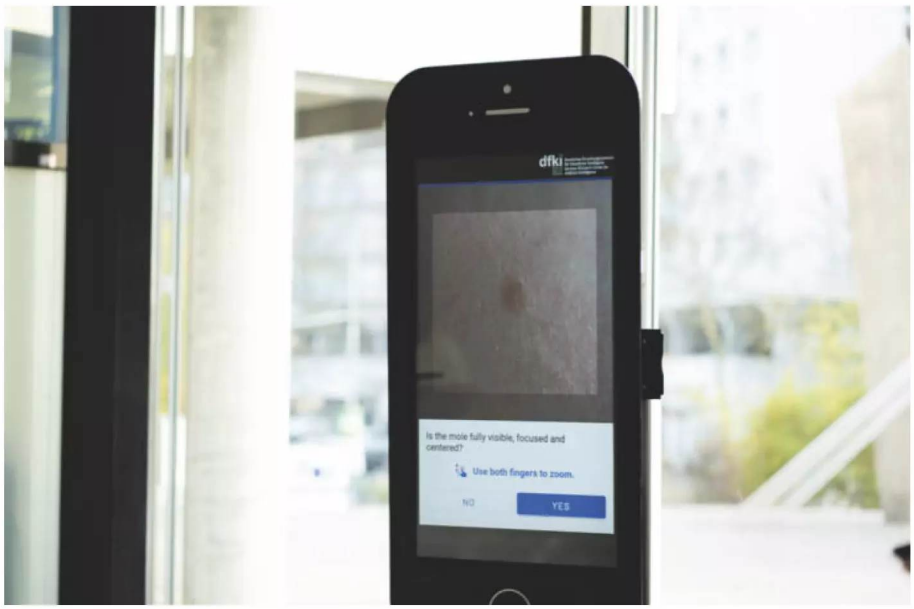
Am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Kaiserslautern hat ein Team um Adriano Lucieri eine Smartphone-App namens SkinDoc für den teledermatologischen Einsatz entwickelt, die ihre KI-Analysen begründen kann. Es ist der Versuch, Fotoaufnahmen von Hautmalen schnell von zu Hause aus einschätzen zu lassen und gleichzeitig die KI-Entscheidungen abzusichern, ohne sich den schon sprichwörtlichen KI-Halluzinationen auszuliefern.

Im Gespräch mit c't berichtet Lucieri, dass er vor zweieinhalb Jahren den Vergleich von Hautscanner-Apps in unserem Magazin gelesen hatte [1]. Zu der Zeit promovierte er zu Methoden der erklärbaren KI. Er erkannte, dass sich gerade die Hautscanner nicht breit durchsetzen können, solange der Anwender nicht weiß, ob ihre Einordnungen gut begründet sind oder nicht.

In Kooperation mit Medizinern der Universitätskliniken in Düsseldorf und Münster entstand zunächst die Anwendung ExAID (Explainable AI for Dermatology), die dermatoskopische Aufnahmen klassifiziert. Das Dermatoskop ist ein Auflichtmikroskop, das es Hautärzten erlaubt, sowohl Male auf der Haut als auch einen Teil von deren Strukturen in tieferen Hautschichten zu betrachten. Entsprechend liefern dermatoskopische Aufnahmen auch mehr Details als einfache Fotos.

Um datenschutzrechtlichen Fragestellungen aus dem Weg zu gehen, trainierte Lucieri das KI-Modell

Bild: DFKI



Im geführten Aufnahmeprozess unterstützen KI-Assistenten dabei, Standards einzuhalten, den betrachteten Leberfleck gut auszuleuchten und ihn beispielsweise auch ins Zentrum der Aufnahme zu rücken.

mit öffentlich zugänglichen Datensätzen wie SkinL2, Derm7pt, PH² und ISIC, von medizinischen Experten annotierte Fotosammlungen, die für die Forschung bereitgestellt worden sind. Er beschränkte sich bei dem Projekt auf die Klassifizierung von Hautkrebsformen. Für das Training des zugrundeliegenden KI-Modells dienten dann standardisierte Aufnahmen in einer hohen vierstelligen Anzahl. Im Vergleich zu einigen kommerziellen Scannern wirkt diese Datenbasis allerdings klein. SkinVision wirbt beispielsweise damit, auf eine Datenbank von 2,9 Millionen Fotos zurückgreifen zu können, „die von unseren Dermatologen bewertet wurden". Über 100.000 dieser Aufnahmen seien schließlich in das Training des KI-Modells eingeflossen.

Konzepte erkannt, Klassifikation begründet

Doch der Ansatz der Forscher ist ein anderer: Sie wollen KI-Modelle schaffen, die ihre Befunde für den Menschen verständlich begründen und belegen können. Gemeinsam mit den beteiligten Medizinern listete Lucieri Erkennungsmerkmale von Hautkrebsarten auf, anerkannte Biomarker. Beispielsweise zeichnen sich viele bösartige Hautmale durch ihre asymmetrische Form aus. Ebenso können unterschiedliche Farben sowie helle und dunkle Bereiche innerhalb eines Hautflecken ein schlechtes Zeichen sein. Verwischte Ränder statt klar definierter Abgrenzungen zur übrigen Haut sind ein weiteres Warnzeichen. Pigmentnetzwerke, also Strukturen, die wie ein auffällig dunkleres Netzwerk aussehen, können typisch oder atypisch ausfallen und so Hinweise für oder gegen Hautkrebs liefern. Das Gleiche gilt für sogenannte Dots and Globules, das sind kleine und verteilte Pünktchen aus dunkleren Pigmenten.

Die Forscher wandten ihre Methoden der erklärbaren KI an und entdeckten im neuronalen Netzwerk ihres Scanners, dass ihr Modell die gesuchten Konzepte auf höheren Schichtebenen tatsächlich anhand der annotierten Trainingsbilder gelernt und manifestiert hatte.

Zu sieben Regeln bauten sie also in ExAID ein Feedback für den Anwender ein. Wenn eine Läsion zum Beispiel auffällig asymmetrisch ist, dann berechnet die Anwendung dazu zunächst einen Konzeptscore. Eine spezielle Heatmap hebt die speziell für dieses Konzept relevanten Bildbereiche hervor, das heißt, die App zeichnet die Ränder deutlich nach. Der zusätzlich ausgegebene Erklärtext weist zudem auf die Asymmetrie hin. Der Vorteil für den Anwender ist, dass er sowohl das Abzeichnen des Randes gut kontrollieren als auch das Konzept Symmetrie versus Asymmetrie leicht nachvollziehen kann. Ähnlich geht es ihm auch mit den anderen Konzepten, die ExAID abcheckt. Auf diese Weise kann der Anwender also die Einschätzungen des Hautscanners mit eigenen Augen überprüfen und eigenständig entscheiden, ob er die Klassifikation für stichhaltig hält oder ob er sie als offensichtliche Fehleinschätzung verwirft.

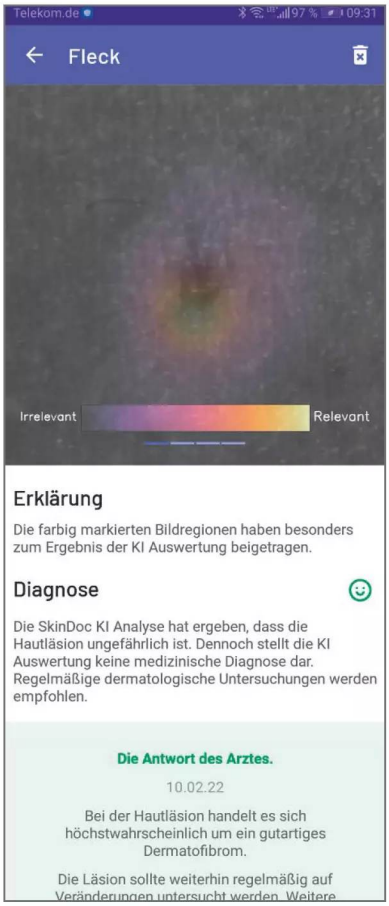
Zusätzlich zu den spezifischen Hautkrebsmerkmalen überprüft das KI-Modell, welche Bildbereiche für die Klassifikation eines Fotos generell entscheidend gewesen sind. Eine sogenannte Heatmap zeigt an, welche Bildpixel sich besonders stark auf das Ergebnis ausgewirkt und letztlich den Ausschlag für die Klassifikation gegeben haben. Auch das ist ein Hinweis, den der Anwender sehr gut überprüfen kann. Weist die Heatmap vor allem auf Bildbereiche hin, auf denen die fotografierte Läsion gar nicht zu sehen ist, dann hat die Klassifikation, ob gutartig oder bösartig, mit dem unter Verdacht stehenden Hautmal nicht viel zu tun.

Erweiterung auf Handyfotos

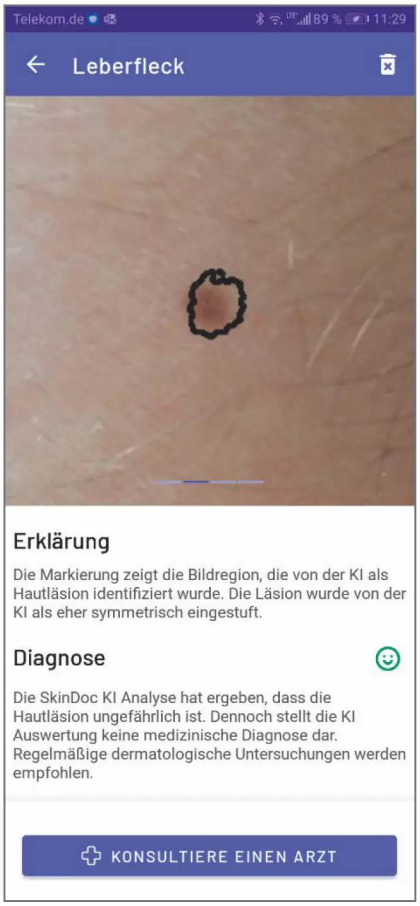
In einem zweiten Schritt entwickelte Lucieri mit SkinDoc eine Smartphone-App, die die Ergebnisse des dermatoskopischen KI-Modells auf eine Anwendung mit einfachen Handyfotos übertragen soll. Um dabei die Qualität der Fotoaufnahmen zu optimieren, geben assistierende KI-Tools bereits während der Aufnahme Feedback oder bewerten Fotos direkt nach dem Auslösen. Beispielsweise erkennen die Tools, ob das Bild unter- oder überbelichtet ist und ob die fokussierte Läsion auch wirklich zentriert aufgenommen wurde. Eine permanente Aufgabe ist es derzeit, die App kontinuierlich an den State of the Art der Kameratechnik anzupassen.

Der zweite Punkt besteht darin, die dermatologischen Konzepte umzusetzen, sodass auch SkinDoc seine Einschätzungen begründen kann. Derzeit kontrolliert diese App automatisiert die Symmetrie einer Läsion und liefert standardmäßig auch eine Heatmap. Die Überprüfung der übrigen Biomarker ist zwar intern implementiert, wird aber von der App noch nicht Konzept für Konzept erklärt.

Zumindest in der jetzigen Form geht SkinDoc ebenso wie die bisherigen Hautscan-Apps den Weg, dass es den Anwender auf die Gefahren einer automatisierten Einschätzung einfacher Smartphone-Fotos



Die Heatmap zeigt, dass im Wesentlichen der betrachtete Leberfleck und sein Umfeld für die KI-Klassifikation entscheidend waren. Aus dem Backend ist in diesem Beispiel bereits eine zusätzliche ärztliche Einschätzung eingetroffen.



Was hat die Entscheidung der KI beeinflusst? Ein KI-Assistent hat die Ränder des Hautmals nachgezeichnet und erklärt daran, dass er dessen Form als eher symmetrisch einstuft. So kann der Anwender diese Entscheidung nachvollziehen.

hinweist. Ein False Positive, also eine unbegründete Warnung vor einem vermeintlichen Hautkrebs, könnte einen Nutzer sehr beunruhigen. Noch schlimmer wären allerdings die Auswirkungen von False Negatives: Ein Patient mit einer schweren Hauterkrankung erhält von seiner Hautscan-App ein beruhigendes Ergebnis und geht daraufhin nicht zu einer dermatologischen Praxis. Um diesen Fall zu vermeiden und etwaigen Schadenersatzforderungen aus dem Weg zu gehen, kommen auch die Forscher derzeit um einen Warnhinweis nicht herum: Wer verdächtige Hautmale an sich entdeckt, sollte eine fachärztliche Praxis aufsuchen.

Teledermatologische Pipeline

Auf lange Sicht strebt Lucieri an, mit SkinDoc eine teledermatologische Pipeline zu etablieren. Zu-

nächst dient die App als Frontend für den Patienten, der damit einen Leberfleck an seinem Körper dokumentieren kann, dazu etwaige Symptome angibt und eine relativ schnelle KI-Erstauswertung bekommt. Diese teilt ihm mit, dass seine Hautläsion eher risikobehaftet ist oder eher weniger riskant. Anhand der dargestellten Konzepte ist auch der Laie in der Lage, selbst zu beurteilen, ob die Ersteinschätzung plausibel erscheint oder ob die KI einen offensichtlichen Fehler gemacht hat.

Auf der anderen Seite gibt es das Backend. Dabei kann es sich um einen zentralen Pool von dermatologischen Fachleuten handeln oder um die dermatologische Praxis, die für den Patienten zuständig ist. An diese Stelle kann SkinDoc seine detaillierteren Einschätzungen senden, beispielsweise einen konkreten Verdacht auf ein Melanom oder ein Basalzellkarzinom. Die zusätzlichen Informationen auch zur KI-Auswertung der erkannten Biomarker können dann assistierend mitgeliefert werden. „Die eigentliche Diagnose wird aber immer von einem Menschen zu stellen sein“, betont Lucieri gegenüber c’t.

Und dann sind da auch noch die Hausarztpraxen. Auch dort haben die Ärzte nur selten ein Dermatoskop zur Hand. Vielmehr gucken sie mit bloßem Auge auf Hautläsionen, die ihnen ihre Patienten ratsuchend zeigen. Auch an dieser Stelle könnte SkinDoc ein Hilfsmittel sein, um Ersteinschätzungen abzusichern und die Kommunikation zwischen Patient, Hausarzt und dermatologischer Praxis zu verbessern und zu beschleunigen.

Menschenzentriert statt KI-hörig

Seit Ende Januar 2025 wird SkinDoc an einem überdimensionalen Smartphone im neuen KI-Innovations- und Qualitätszentrum (IQZ) in den Räumen des Deutschen Technikmuseums in Berlin als laufendes Forschungsprojekt öffentlich präsentiert. Dort ist die App ein Bestandteil der nationalen Initiative „Mission KI“. Das DFKI fokussiert darin medizinische KI-Anwendungen, deren Vertrauenswürdigkeit sich überprüfen lässt. Dementsprechend steht für Lucieri und seine Kollegen nicht im Vordergrund, möglichst viele Trainingsbilder zu sammeln, um damit die Diagnosegenauigkeit des zugrundeliegenden KI-Modells zu perfektionieren. Zusätzliche Trainingsdaten sollen stattdessen dazu dienen, die konzeptbasierten Erklärungen zu verbessern. Wichtig ist ihnen die Menschenzentriertheit ihrer App. Mit den Methoden der erklärbaren KI wollen sie die Entscheidungskonzepte ihres Modells weiter genau analysieren und mit den medizinischen Erkenntnissen abgleichen.

Zudem wollen sie die Erklärmethoden verbessern, sodass der Anwender mit wenigen Blicken versteht, worauf die KI-Entscheidungen beruhen. Nur so kann er stichhaltige Klassifikationen von Fehleinschätzungen unterscheiden. Und nur so wird es möglich, in Zukunft einen Hautkrebsscanner aufs Smartphone zu bringen, der dem Patienten oder der Patientin verständlich und nachvollziehbar mehr raten kann als die versicherungstechnische Floskel: „Wenn Ihnen ein Hautmal verdächtig vorkommt, gehen Sie damit am besten zu Ihrer dermatologischen Praxis.“ (agr) ct

Literatur

[1] Arne Grävemeyer, Gefährliche Male?, Hautscan-Apps mit KI im Vergleich, c’t21/2022, S. 124

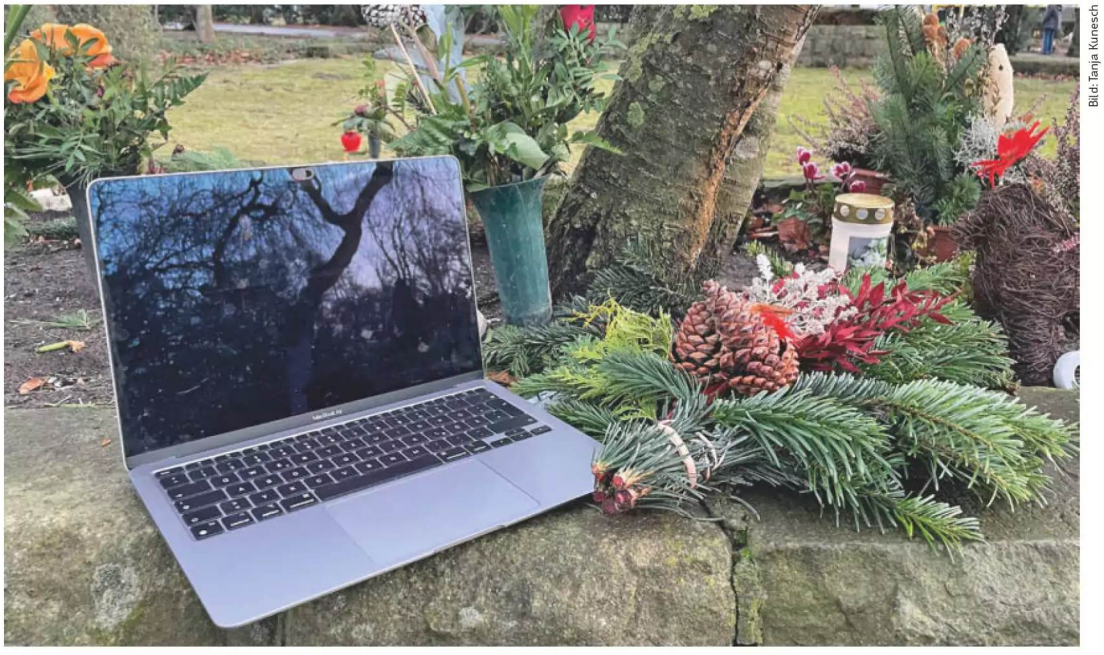


Bild: Tanja Kunesch

Digital unsterblich durch KI

In der digitalen Welt endet das Leben nicht mehr zwingend mit dem Tod. Verwaiste Facebook-Profile wirken so, als seien Verstorbene weiterhin erreichbar. QR-Codes auf Gräbern führen zu digitalen Trauerräumen und Avatare von Verstorbenen begleiten die Hinterbliebenen. Aber wie hilfreich ist die Begleitung der Trauer durch Technik? Schadet sie womöglich mehr, als sie nutzt?

Von **Tanja Kunesch**

Längst sind Menschen imitierende Roboter keine Neuheit mehr – zumindest in Film und Fernsehen. Vielleicht waren wir 2013 noch schockiert, als in einer Folge der dystopischen TV-Serie „Black Mirror“ die Protagonistin Martha ihren verstorbenen Freund und Vater ihrer Tochter mithilfe einer KI nachbaut, weil sie mit seinem Verlust nicht umgehen kann.

Was als ethisches Gedankenspiel begann, schleicht sich langsam in die Realität. Seit etwa einem Jahrzehnt gibt es immer wieder Berichte über verschiedene Start-ups, die ein digitales Weiterleben versprechen. „Digital Afterlife Industry“ (DAI) nennt sich die Branche, die Technik benutzt, um sich mit Tod und Trauer auseinanderzusetzen. Einige Anbieter wollen Hinterbliebenen Räume zum Erinnern bieten.

Andere nutzen Daten von Verstorbenen, um Avatare zu erstellen, mit denen sich Trauernde unterhalten können. Sie bieten Chats oder gar virtuelle Realitäten. Manche geben nur bekannte Textschnipsel wieder, andere nutzen künstliche Intelligenz, um den Verstorbenen so glaubwürdig wie möglich zu imitieren.

Und wieso auch nicht? Wer würde nicht gerne noch einmal mit einem geliebten Menschen sprechen, den er oder sie verloren hat? Oder enden Versuche, mithilfe einer Maschine der Trauer zu entkommen, zwangsläufig wie bei Martha – unfähig, mit der Vergangenheit abzuschließen?

Unter anderem mit diesen Fragen befasst sich Matthias Meitzler seit Jahren. Er ist wissenschaftlicher Mitarbeiter an der Universität Tübingen, studierter Soziologe und Psychoanalytiker. Besonders intensiv hat sich Meitzler mit Technikphilosophie auseinandergesetzt. Unter anderem hat er untersucht, welche technischen Anwendungen es für Tod und Trauer gibt und was sie für Auswirkungen auf die Trauerkultur in der Gesellschaft haben: Wann können sie Menschen bei ihrer Trauer helfen und wann schaden sie?

Um das herauszufinden, hat der Soziologe an dem Projekt Edilife mitgewirkt, das von 2022 bis 2024 lief und seine Ergebnisse bereits veröffentlicht hat (Studie auf ct.de/w8bz). Das kooperative Forschungsprojekt der Universität Tübingen und des Fraunhofer SIT (Darmstadt) untersucht, welche virtuellen Existenzen es nach dem Tod geben kann und wie sie unsere Gesellschaft beeinflussen. Eine entscheidende Frage ist, ob ein Avatar die Verlustbewältigung für Hinterbliebene erleichtert oder zusätzlich erschwert.

Ein digitales Abbild

Ein Avatar ist eine virtuelle Existenz nach dem Tod und meint eine Software, die Kommunikation von Verstorbenen nachahmt, in manchen Fällen sogar inklusive ihrer Stimme und ihres Aussehens. Dafür benötigt sie große Datenmengen wie Chats, Sprachnachrichten oder Videoaufnahmen, die der Mensch zu Lebzeiten produziert hat.

Tatsächlich sind viele der KI-Angebote noch gar nicht wirklich auf dem Markt, berichtet Meitzler. „Aber es ist schon manches möglich. Und wenn wir uns anschauen, was in den letzten Jahren im Bereich der KI passiert ist, wird hier wohl noch einiges kommen“, sagt er.

Deutsche Firmen gibt es keine. Das habe vor allem wirtschaftliche Gründe. „Die Top Player sind

Bild: Matthias Meitzler

Matthias Meitzler ist Soziologe und Psychoanalytiker der Universität Tübingen. Er fragt sich: Wann kann Trauertechnologie Hinterbliebenen helfen und wann schadet sie?

in den USA und in Ostasien“, so Meitzler. In den USA gibt es etwa „You Only Virtual“ (erwähnte Websites auf ct.de/w8bz). Das Unternehmen wirbt auf seiner Homepage mit den Worten „Pioneering the technology of never saying goodbye“. Kunden können dort eine digitale Version ihrer selbst kreieren, eine „Versona“, die nach ihrem Ableben weiter mit den Angehörigen kommuniziert – vor allem über Chats.

Oder Eter9: Auf seiner Website wirbt das Unternehmen sogar mit einer Rezension, die auf die Serie Black Mirror anspielt: „Black Mirror predictions that disturbingly came true“. Gemeint ist damit das „digital immortality feature“, das dem Kunden erlaubt, einen digitalen Zwilling von sich zu erschaffen, der sogar online weiter in dessen Sinne Kommentare postet und in der Netzwelt aktiv bleibt [1]. Beide Angebote können auch Kunden in Deutschland nutzen.

Meitzler unterscheidet die bisherigen Anwendungen nach ihrem Output. So geben einige das eingespeiste Material, etwa in Form von Sprachnachrichten, unverändert aus. „Meine Hinterbliebenen können Fragen an die App richten und diese sucht mithilfe von KI die entsprechenden Ausschnitte raus

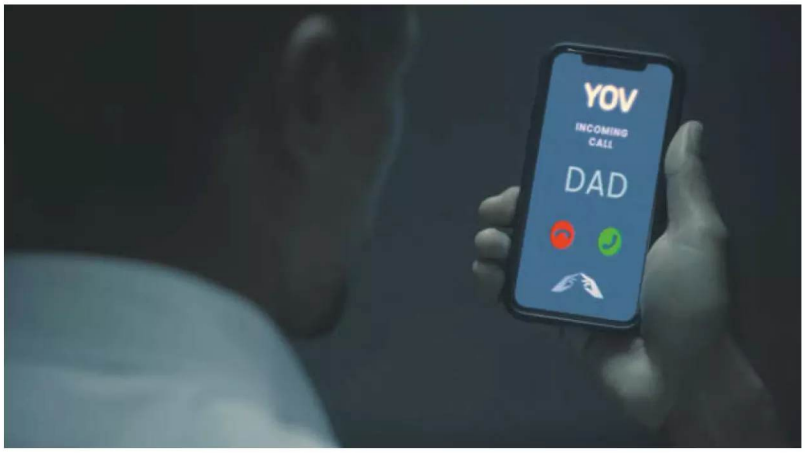


Bild: myyov.com

Das Unternehmen „You Only Virtual“ hilft Kunden, digitale Versionen ihrer selbst zu kreieren, sogenannte Versonas.

und spielt sie ab“, sagt der Soziologe. Andere nutzen das Material nur als Ausgang und ändern es: Ein Chatbot unterhält sich schriftlich mit dem Nutzer, ein Avatar kommuniziert zudem visuell.

Allerdings kommt die Studie Edilife zu dem Schluss, dass die Anwendungen in ihrer Nutzung noch sehr eingeschränkt sind. Zukünftig werden Avatare wohl in digital erweiterte Welten (Augmented Reality, AR) integriert und sich zunehmend zu virtuellen Realitäten entwickeln. Heutige AR-Anwendungen sind bisher nicht in der Lage, eine glaubhafte Kommunikation zwischen Audio-Video-Avataren und anwendenden Personen abzubilden, heißt es weiter. Die Dynamik der menschlichen Kommunikation könne bisher von den Avataren nur ungenügend imitiert werden.

Trauer als Geschäftsmodell

Doch es muss nicht immer gleich eine KI sein. So gibt es etwa auf Facebook die Möglichkeit, Profile von Verstorbenen in einen Gedenkmodus zu versetzen, um sie noch eine Zeitlang als Ort des Erinnerns zu erhalten. Vereinzelt finden Besucher auf Friedhöfen QR-Codes auf Grabsteinen, die zu einem digitalen Erinnerungsraum für die Verstorbenen führen. Oder Hinterbliebene können sich eine App herunterladen, die ihnen in der Trauerzeit beistehen soll, wie etwa Grievy.

2021 hat Nele Stadtbäumer mit zwei weiteren Entwicklern die App als eine „vertrauenswürdige, barrierearme und sofortige digitale Lösung für den Trauerprozess“ geschaffen. Darin finden Hinterbliebene nicht nur Checklisten und Hilfe, um die Beisetzung zu organisieren. Vor allem will Grievy seine Nutzer bei ihrer Trauer begleiten und stellt mehr als 150 Übungen bereit, die laut Website von Psychologen entwickelt wurden und wissenschaftlich validiert sind.

Meitzler hat sich in seiner Forschung auch mit dieser App befasst. Er sieht darin ein niedrigschwelliges Angebot, das durch seine digitale Form manchen Menschen die Hemmung nimmt, sich mit ihrer Trauer auseinanderzusetzen. „Außerdem ist es heutzutage oft nicht so leicht, auf die Schnelle eine geeignete psychologische Betreuung zu finden“, fügt der Soziologe hinzu.

Bild: Matthias Meitzler/Thorsten Benkel

QR-Codes auf Grabsteinen führen zu digitalen Erinnerungsräumen, die Hinterbliebenen helfen sollen, der Verstorbenen zu gedenken.

Bild: youtube.com/@MBClife

Mithilfe einer VR-Brille verabschiedete sich eine Mutter in Südkorea von ihrer verstorbenen Tochter.

Wann kann KI bei Trauer helfen?

So unterschiedlich und innovativ diese Anwendungen sind, bleibt die Frage, ob sie Trauernden wirklich helfen können. „Wir können da tatsächlich noch auf keine große empirische Datenbasis zugreifen. Insofern ist es schwierig, pauschal zu sagen, ob es in der Trauer hilft oder vielleicht sogar schadet“, sagt Meitzler. Er plädiert dafür, digitale Trauerunterstützung nicht von vornherein zu verteufeln, sondern zu differenzieren und zu begleiten. Wer ist wie gestorben? Wann nutzen Hinterbliebene das Angebot und wie oft?

Manche Fälle erscheinen wenig riskant und haben sogar eine spielerische Komponente. Wie etwa ein Fall aus Südkorea, der 2016 weltweite Aufmerksamkeit erregt hatte (Link zum Video auf ct.de/w8bz). Eine Mutter konnte ihrer Tochter, die mit sieben Jahren verstorben war, mithilfe einer VR-Brille noch einmal begegnen. Das habe ihr geholfen, Abschied zu nehmen. Gleichzeitig war es eine einmalige Begegnung und keine dauerhafte Begleitung, wie Chatbots sie anbieten würden.

Risiken eines Avatars

Und das ist genau einer der Risikofaktoren: Was, wenn Nutzende in eine Parallelwelt abdriften und den Bezug zur Realität verlieren, weil sie den Verlust einer Person nicht akzeptieren können? Menschen neigen laut Meitzler dazu, unbeseelten Dingen wie einer Maschine Lebendigkeit zuzuschreiben. „Und jetzt hat man einen Avatar vor sich, der so aussieht wie die geliebte Person, der genauso spricht. Der Avatar könnte dann eine gewisse Suggestivkraft entfalten, die mich glauben lässt: Das muss dieser Mensch sein.“ Gerade ein dramatischer Verlust wie ein Suizid eines jungen Menschen lasse viele Fragen offen. „Da könnte ich mir vorstellen, dass sich manche Hinterbliebene an einen Avatar richten – der dann tatsächlich antwortet“, so Meitzler.

Ein Blick hinter die Kulissen offenbart eine weitere Problematik: Die Avatare werden für gewöhnlich von kommerziellen Unternehmen geschaffen, die Menschen mit ihren Angeboten an sich binden wollen. Allein im Jahr 2023 starben in Deutschland 1,03 Millionen Menschen. Wirtschaftlich gesehen ist die Branche damit ein sicherer Markt: Der Tod ist schließlich garantiert und wird immer Hinterbliebene und damit potenzielle Kunden generieren. „Aber was, wenn ich das Angebot für mich persönlich nicht mehr benötige und der Avatar mich dazu drängt, das Abo zu verlängern? Er könnte dann beispielsweise versuchen, mir ein schlechtes Gewissen einzureden, indem er sagt: Bitte lass mich kein zweites Mal sterben!“, sagt der Soziologe.

Der Avatar kann so zu einer Droge werden, zur Sucht. Auch hier zieht Meitzler Vergleiche zu früheren Entwicklungen. So habe die Fotografie bei ihrem Aufkommen im 19. Jahrhundert ebenfalls starke Skepsis hervorgerufen: Menschen würden ihrer Seele beraubt werden und es sei gruselig, die Gesichter der Toten im Wohnzimmer hängen zu haben. „Passiert mit der KI und den Avataren nicht Ähnliches, nur eben auf einem anderen technischen Level? Und können wir in Zukunft, trotz aller Risiken, einen

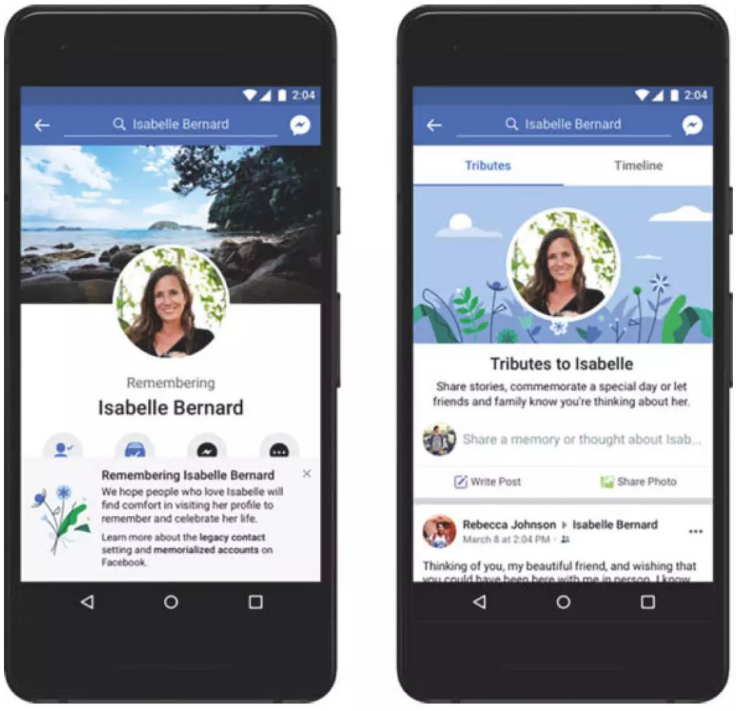


Bild: Meta

Hinterbliebene können Facebook-Profile von Verstorbenen in den Gedenkzustand versetzen, um sie zu ehren.

selbstbestimmten und verantwortungsbewussten Umgang lernen, der letzten Endes mehr nützt als schadet?“, fragt der Soziologe.

Für Hinterbliebene, denen es gelingt, zu verstehen, dass der Avatar nicht die eigentliche Person ist, sieht Meitzler darin die Möglichkeit, sich mit einem abgeschlossenen Leben und seiner eigenen Trauer auseinanderzusetzen. „So was kann durchaus Reflexionsprozesse auslösen, die einem bei der Bewältigung des Verlustes helfen“, sagt er. Generell gebe es hier wohl kein Richtig und Falsch, sondern für individuelle Personen Hilfreiches und weniger Hilfreiches.

Einschätzung einer Trauerbegleiterin

Dem individuellen Charakter von Trauerprozessen stimmt Christine Kempkes zu. Sie ist Trauerbegleiterin und bildet selbst Trauerbegleitende nach den Standards des Bundesverbands Trauerbegleitung e.V. aus. „Trauer ist erst einmal das ganz normale Gefühl, wenn wir jemanden oder etwas Wertvolles verloren haben“, sagt sie. „Dabei dürfen Menschen zwei Schritte vor und einen zurückgehen. Das ist alles ein gesunder Rahmen, wenn ich insgesamt merke, sie trauern auf das Leben zu.“

Mit einer KI zu chatten, sieht Kempkes allerdings extrem kritisch: „Eine Aufgabe, die wir als Trauernde bewältigen müssen, ist es, die Realität, so wie sie jetzt ist, zu akzeptieren. Wirklich zu verstehen, dass der Mensch tot ist. Eine KI gaukelt mir vor, dass es diesen Menschen doch noch gibt.“ So versuche sie, etwas fortzusetzen, was nicht mehr existiert. Solange Menschen den Verlust nicht akzeptieren, könne kein heilsamer Trauerprozess beginnen.

Wichtiger sei es, eine neue Verbindung zu dem Verstorbenen zu finden, um der Person nahe zu sein, vielleicht durch die Erinnerung an den früheren gemeinsamen Urlaubsort am Meer. Wer seinen Liebsten den Abschied leichter machen will, der kann schon zu Lebzeiten Audioaufnahmen von Gesprächen machen, Briefe schreiben und Botschaften hinterlassen. „Das ist handgeschrieben von der

echten Person und kann später eine heilsame Erinnerung sein.“

Natürlich können Menschen auch ohne eine KI in ihrer Trauer verharren. Das sieht Kempkes mitunter darin begründet, dass unsere Gesellschaft verlernt hat, zu trauern. „Wir wollen funktionieren. Wir wollen ein glatt poliertes, schönes Leben haben. Spätestens nach drei Monaten fragen wir den Trauernden: Bist du denn immer noch traurig? Dabei wäre es doch total kurios, wenn ein Mensch, den ich sehr geliebt habe, verstorben ist und ich nach drei Monaten nicht mehr todtraurig bin.“

Genau diese Lücke im Umgang mit den Toten nutzen Digital-Afterlife-Unternehmen, indem sie einen vermeintlichen Ausweg bieten. Helfen werden diese Angebote nicht, da ist sich die Trauerbegleiterin sicher. „Es ist und bleibt ja eine Maschine. Und Maschinen begleiten keine Menschen. Menschen begleiten Menschen.“

Das hat Kempkes in den vergangenen Jahren immer wieder erlebt. Sie bekommt Anfragen von Menschen, die oft bereits ein digitales Angebot ausprobiert haben, wie etwa die App Grievy. Viele wollen am Ende doch einfach mit einem anderen Menschen sprechen und landen dann bei ihr oder ihren Kollegen.

Dabei findet die Trauerbegleiterin solche digitalen Anlaufstellen wichtig, gerade in ländlichen Gebieten, wo es nicht so viele Trauergruppen vor Ort gibt. Bestimmte Angebote vermitteln sogar Trauerfreundschaften, wie die Seite trosthelden.de. Personen in vergleichbaren Situationen zu finden und sich auszutauschen, kann ebenso helfen.

Gerade diese Delokalisierung von Trauer ist einer der größten Wandel, die die Technik auf unsere Trauerkultur bereits eingeleitet hat. Trauer muss nicht mehr dort sein, wo der Körper beigesetzt wird. Jeder mit einem Internetzugang kann im Prinzip mittrauern. Und mitgestalten.

Datenschutz: Einen absoluten Schutz gibt es nicht

Diese Flexibilität bringt wiederum ihre eigenen Risiken mit sich. Je mehr Menschen im Internet von sich preisgeben, desto mehr sensible Daten kursieren, seien es digitale Trauerräume, Chats oder Avatare.

Wie vielfältig die Risiken gerade bei den KI-gestützten Anwendungen sind, erklärt Annika Selzer. Sie ist Leiterin der Forschungsabteilung „IT Law and Interdisciplinary Privacy Research“ am Fraunhofer SIT und koordiniert den Forschungsbereich „Legal Aspects of Privacy and IT-Security“ im Nationalen Forschungszentrum für Angewandte Cybersicherheit Athene. Sie warnt: Verstorbene könnten etwa ungewollt als Avatar abgebildet werden, der Avatar könnte vertrauliche Informationen weitergeben, sein Wesen verändern oder einfach ungewollt gelöscht werden.

Bild: Eve Sundermann-Biesen

Christine Kempkes ist Trauerbegleiterin. Helfen werden die Angebote der Digital-Afterlife-Unternehmen nicht, da ist sie sich sicher.

Zwar gebe es in Deutschland einen postmortalen Persönlichkeitsschutz, der zum Beispiel vor Erniedrigungen oder diffamierenden Darstellungen durch Dritte schützen solle – analog zu Deepfakes von echten Personen in Pornos. Auch der Name und das Bildnis seien grundsätzlich abgesichert. Allerdings könne es bei Allerweltsnamen wie Hans Schmidt schwierig werden oder wenn eine Abbildung nicht sonderlich originalgetreu sei.

Zudem gelten diese Achtungsansprüche nicht auf unbestimmte Zeit. „Das geltende Recht geht davon aus, dass dieser Schutz genauso verblasst wie eben auch die Erinnerung an einen Menschen nach dessen Tod. Deswegen nimmt der Schutz aus diesen Vorschriften mit der Zeit immer mehr ab“,

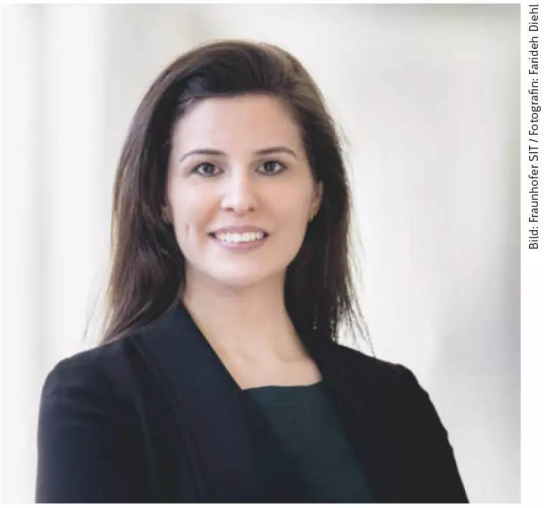


Bild: Fraunhofer SIT / Fotografin: Fariden Diehl

Vielen Menschen fehle das Bewusstsein für ihr digitales Erbe, sagt die Datenschutzexpertin Annika Selzer.

sagt Selzer. Wie lange es bis zum Erlöschen dauert, sei individuell.

Wer sichergehen will, kann sich eigentlich nur schützen, indem er seinen Willen schriftlich in einer letztwilligen Verfügung festhält [2]. Die lässt sich zusätzlich mit Auflagen kombinieren, die dafür sorgen, dass der letzte Wille umgesetzt wird, wenn nötig auch mit einem Testamentsvollstrecker.

Doch einen absoluten Schutz gebe es nicht, sagt Selzer. Selbst wenn jemand einen Testamentsvollstrecker einsetzt, könne ein Dritter die Daten, die über den Verstorbenen kursieren, im stillen Kämmerlein benutzen.

Dabei könnten Dienstanbieter mehr Schutz bieten. Allerdings ist das gerade für international agierende Unternehmen mit viel Aufwand verbunden. Denn wenn sie ihren Nutzern individuelle Einstellungen ermöglichen, müssen diese mit dem jeweiligen nationalen Erbrecht im Einklang stehen. „Andererseits müssen zum Beispiel Social-Media-Plattformen in ihren Angeboten auch etliche nationale Vorschriften, etwa für den Datenschutz, berücksichtigen. Sich mit unterschiedlichen nationalen Rechtsrahmen auseinanderzusetzen, wäre also für diese grundsätzlich nichts Neues“, sagt Selzer.

Prinzipiell findet die Datenschutzexpertin den postmortalen Schutz in Deutschland recht gut – wenn wir ihn denn wahrnehmen. Das Problem sei eher das fehlende Bewusstsein dafür. „Durch diesen weiterlebenden Avatar hört das Leben auf einmal nicht mit dem Tod auf. Und das sind wir nicht gewohnt, sind uns der Folgen noch nicht gut genug bewusst. Aber auch abseits von weiterlebenden Avataren fehlt vielfach noch das Bewusstsein für das digitale Erbe – viele hinterlassen ihren Erben einen digitalen Friedhof an Daten und an Plattform-Accounts.“ [3]

Dass mit KI-Avataren das Risiko steigt, dass sensible Daten geklaut werden, sei kein reines Problem der Digital Afterlife Industry: „Je mehr wir unsere Stimme oder Videos verfügbar machen, umso wahrscheinlicher ist es, dass diese Aufzeichnungen genutzt werden können, um damit Schabernack zu treiben.“ Doch generell davon abraten will Selzer nicht. Wer sich dafür entscheidet, einen KI-gestützten Digital-Afterlife-Dienst zu nutzen, sollte das bewusst tun und sich vorab umfangreich informieren: Wie lange gibt es den Dienst schon? Inwieweit hat er festgelegt, wie er sensible Daten und somit auch die Menschen schützt, zu denen diese Daten gehören?

Ausblick auf die Zukunft

Sorgen darüber, dass bald alle mit den Avataren von Verstorbenen auf dem Sofa sitzen, machen sich die wenigsten. Heidi Müller ist Trauerberaterin und Mitbegründerin der Europäischen Trauerkonferenz, die im November 2024 in Dublin stattfand. Dort war KI in der Trauer nur ein Randthema. „Wir haben im Moment andere Themen, wie Trauer nach Desaster, Trauer in Kriegszeiten oder klimabezogene Trauer“, sagt Müller.

Der Soziologe Meitzler rechnet allerdings damit, dass die Nutzerzahlen steigen werden. Eine Bitcom Studie von 2023 ergab bereits, dass sich ein Drittel der Internetnutzer in Deutschland wünscht, dass ihre Profile in sozialen Netzwerken nach ihrem Ableben fortbestehen. Natürlich gebe es noch gewisse Berührungsängste, das sagen Meitzlers Befragte immer wieder. Das liege aber daran, dass Menschen, die vermehrt mit Sterben, Tod und Trauer zu tun haben, statistisch meist älter sind und keine Digital Natives. „Gerade junge Menschen sind es gewohnt, online nach Hilfe zu suchen und auch über sehr private Dinge zu kommunizieren“, sagt er. Und die würden schließlich irgendwann älter. „Das ist durchaus ein Zukunftsthema.“ (spa) **ct**

Literatur

[1] Arne Grävemeyer, Die Geister, die ich rief, Künstlich intelligente Avatare lassen Tote auferstehen, c’t 17/2019, S. 136

[2] Holger Bleich, Prince Leslie Kwarfo Krow und Marva Pirweyssian, Nachlass mit Recht, Juristische Grundlagen rund ums digitale Erbe, c’t 15/2024, S. 72

[3] Holger Bleich und Dorothee Wiegand, Kontenklärung, Tipps zum Verwalten eines digitalen Nachlasses, c’t 15/2024, S. 64

Studien und erwähnte Websites:

ct.de/w8bz

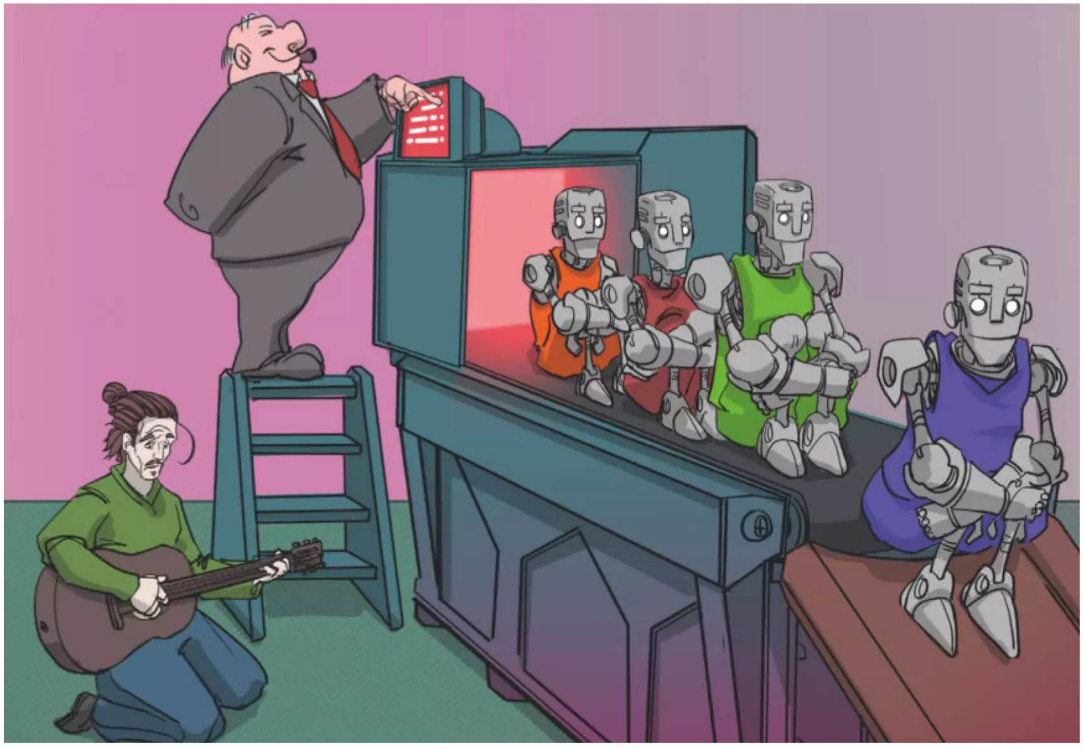
9. Dezember

c't WEBINAR

Wissenschaftlich Schreiben mit KI-Unterstützung

Jetzt informieren:

heise-academy.de/webinare/wissenschaftlich-schreiben



(Bild: Thorsten Hübner)

KI krempelt die Musikindustrie um

KI-Generatoren wie Suno erzeugen Songs ohne Komponisten und Künstler. Die Folgen für die Musikindustrie sind dramatisch. Musiker stehen vor der Frage: Was bleibt, wenn Maschinen ihre Songs produzieren?

Von **Hartmut Gieselmann**

Wer verstehen will, was KI-Generatoren wie Suno oder Udio mit der Musikindustrie anrichten, sollte einen alten Philosophen befragen. Vor über hundertfünfzig Jahren analysierte Karl Marx den kapitalistischen Produktionsprozess – und prognostizierte dessen Untergang. Von künstlicher Intelligenz war damals keine Rede, wohl aber von vollautomatischen Fabriken, die ohne menschliches Zutun Waren am Fließband herstellen.

Betrachtet man die Entwicklung der Musikindustrie – von der Aufzeichnung auf Schellack über die Digitalisierung auf CD, von Streamingdiensten bis

hin zu KI-Generatoren –, scheint dieser Punkt bald erreicht. Auf Knopfdruck generierte KI-Musik markiert jedoch nicht nur einen technischen Fortschritt, sondern auch eine Krise: die der Künstler und die der Musik als konsumierbare Ware.

In diesem Artikel zeichnen wir die Etappen dieser Entwicklung nach und werfen einen Blick in die Zukunft: Was kommt nach den KI-Diensten und wie können Musiker diesen Prozess überleben?

Vom Gebrauchs-, Tausch- und Mehrwert der Musik

Um die Folgen der KI-Musikproduktion zu verstehen, lohnt der Blick auf drei zentrale Begriffe aus der ökonomischen Theorie von Karl Marx: Gebrauchswert, Tauschwert und Mehrwert. Sie lassen sich auch auf Musikstücke anwenden und zeigen, warum KI eine Zäsur in der Produktion darstellt [1].

Der Gebrauchswert eines Musikstücks beschreibt seinen konkreten Nutzen für den Hörer: Ob es seine Stimmung hebt, ihn zum Tanzen animiert oder seine Langeweile vertreibt. Musik wird zuweilen auch gezielt eingesetzt, etwa um Panik in Fahrstühlen zu unterdrücken oder den Kauf von Getränken in Kneipen anzukurbeln. Man nennt sie dann auch „Muzak" – benannt nach einer US-Firma, die seit den 1950er-Jahren zum weltgrößten Hersteller solcher Hintergrundmusik aufstieg. Der Gebrauchswert von Musik lässt sich zudem technisch steigern, etwa indem man den Zugang zur Musik erleichtert oder die Abspielqualität verbessert.

> ***„Es erscheint hier also direkt die bestimmte Arbeitsweise übertragen von dem Arbeiter auf das Kapital in der Form der Maschine und durch diese Transposition sein eignes Arbeitsvermögen entwertet. Daher der Kampf der Arbeiter gegen die Maschine. Was Tätigkeit des lebendigen Arbeiters war, wird Tätigkeit der Maschine."***
>
> *Karl Marx, Grundrisse der Kritik der politischen Ökonomie, MEW Bd. 42, S. 600*

Davon getrennt bemisst der Tauschwert einer Musik, wie viel Zuhörer bereit sind zu bezahlen, um ein bestimmtes Stück zu hören oder eine Platte zu besitzen. Das können Unsummen sein, wenn sie einen Popstar aus der ersten Reihe im Konzert hören wollen – oder gar nichts, wenn sie einen kostenlosen Stream empfangen. Der Tauschwert hängt nicht vom tatsächlichen Nutzen eines Stücks ab, sondern nur von der Beziehung zwischen Anbieter und Käufer auf dem Markt. Werden beispielsweise mehr Produkte angeboten als nachgefragt, dann sinkt der Tauschwert. Der Gebrauchswert bleibt jedoch unverändert.

Der Mehrwert misst schließlich die Differenz zwischen dem Tauschwert, den Künstler und Techniker durch ihre Arbeit schaffen, und dem Lohn, den sie dafür erhalten. Entsteht ein Song durch die Arbeit von Komponisten, Musikern und Toningenieuren und bringt er am Markt mehr ein, als diese dafür bezahlt bekommen, dann eignet sich das Label oder der Plattformbetreiber diesen Mehrwert an. Laut Marx kann er nur aus menschlicher Arbeit gezogen werden, nicht aber aus besseren Maschinen oder einem höheren Grad der Automatisierung.

Vom Mehrwert zur Rente

Dieser Mehrwert war bislang *die* Profitquelle der Musikindustrie. Er ist die zentrale Kategorie, um zu bewerten, was mit der Rolle der Musikschaffenden geschieht, wenn KI-Systeme den Produktionsprozess übernehmen. Auch die US-Firma Muzak entlohnte Musiker und Komponisten und profitierte von dem durch deren Arbeit geschaffenen Mehrwert. Wenn künftig solche Musik jedoch aus einem KI-Generator kommt, entsteht kein Mehrwert mehr. Die KI selbst erhält keinen Lohn. Stattdessen zahlen Nutzer der KI-Plattformen den Betreibern eine Rente in Form von Abo- oder Lizenzgebühren.

Marx bezeichnet diese Rente als ein Einkommen, das man nicht bekommt, weil man mit menschlicher Arbeit etwas produziert, sondern weil man etwas besitzt, zum Beispiel Land, Maschinen oder in diesem Fall einen KI-Dienst. Es findet keine neue Wertschöpfung statt, sondern das bereits von Musikern und Technikern in der Vergangenheit Erarbeitete wird lediglich algorithmisch verwertet. Doch das Rentnerglück der Konzerne währt nicht ewig: Ohne weiteren menschlichen Input kommt die musikalische Entwicklung langfristig zum Stillstand. Denn auch KI-Generatoren sind kein Perpetuum mobile.

Die Betreiber der KI-Plattformen eignen sich aber nicht nur die vergangene Arbeit von Musikern und Bands an, mit der sie ihre Algorithmen trainieren, sondern auch den Input der Nutzer ihrer Dienste. Diese laden neue Musikideen, Texte und Demo-

schnipsel hoch und bewerten den Output der KI-Generatoren, ohne dafür bezahlt zu werden. So entsteht eine sogenannte extraktive Plattformwirtschaft, die mit den Inhalten und Daten der Nutzer Geld verdient, ohne diese dafür zu entlohnen oder am Gewinn zu beteiligen.

Das sorgt zwar kurzfristig für enorme Profite, weil die bisherigen Lohnkosten für Künstler und Tontechniker entfallen, hat aber langfristig den Nachteil, dass sich die Musik kaum noch weiterentwickelt. Denn dazu müssten echte professionelle Musiker neue Ideen in das System einspeisen. Und dafür muss man sie entlohnen. Andernfalls droht ein kultureller Stillstand und das Geschäftsmodell der Musikindustrie zu kollabieren.

Dies ist jedoch kein zwangsläufiger Prozess. In der Geschichte der Musikindustrie gab es immer wieder Krisen durch technische Umbrüche und Versuche, darauf mit neuen Verwertungsmodellen zu reagieren. Dazu ein Blick in die Vergangenheit.

Von der Schellackplatte zur CD

Mit der Einführung der Schellackplatte Ende des 19. Jahrhunderts wurde Musik erstmals dauerhaft speicher- und verkaufbar. Das steigerte ihren Gebrauchswert, da man nicht mehr in ein Konzert gehen musste, um flüchtige Aufführungen zu hören, sondern Musik im Alltag von einem Grammophon abspielen konnte – immer und immer wieder. Der Tauschwert spiegelte den Preis der Schellackplatten wider, die in immer größeren Stückzahlen verkauft wurden. Um diese zu produzieren, nahmen die Plattenlabel Komponisten und Musiker unter Vertrag und stellten Toningenieure sowie Arbeiter in den Presswerken ein. Das Label bezahlte ihnen dafür einen Lohn und profitierte vom Mehrwert, den es durch den Verkauf der Schallplatten erzielte.

Ab den 1950er-Jahren verbreiteten sich Tonbandgeräte und später die Compact-Kassette, mit denen die Konsumenten selbst Musik aufzeichnen und kopieren konnten. Mitschnitte aus dem Radio und der Tausch mit Freunden erhöhten abermals den Gebrauchswert der Musik. Der Tauschwert wurde jedoch instabil. Kopien unterliefen die Verkäufe der bisherigen Tonträger und reduzierten den Profit der Plattenlabels. Diese wehrten sich juristisch und suchten nach neuen Vertriebskonzepten.

Den nächsten Push löste die Compact Disc Anfang der 80er-Jahre aus. Sie war das erste digitale Medium, das sich verlustfrei vervielfältigen und abspielen ließ. Die kleinen Plastikscheiben konnten billiger produziert werden als Vinyl-Schallplatten, boten eine weit bessere Qualität als Musik-Kassetten und erlaubten, Titel direkt anzuspringen. Dadurch stieg nicht nur der Gebrauchswert, sondern auch der Tauschwert, weil Kunden bereit waren, höhere Preise zu bezahlen. Kostete eine LP damals durchschnittlich 15 bis 20 Mark, so waren es für eine CD rund 30 Mark. Das steigerte den Mehrwert für die Labels enorm, die zudem alte Aufnahmen auf CD erneut verkaufen konnten. Es bildete sich ein Oligopol weniger großer Verlage, die bald den gesamten Musikmarkt weltweit kontrollierten.

Entmaterialisierung der 90er

Doch der Einfluss der Musikverlage begann zu bröckeln, als Ende der 90er-Jahre die digitalen Musikdaten im MP3-Format so stark komprimiert werden konnten, dass sie sich wenig später massenhaft über das Internet tauschen und auf Festplatten sammeln ließen. Mit dieser Entmaterialisierung der Musik explodierte ihr Gebrauchswert: Millionen von Songs wurden quasi kostenlos jederzeit verfügbar. Dadurch kollabierte jedoch der Tauschwert und damit auch der Mehrwert, den die Labels abschöpfen konnten.

> ***„Alles wird Muzak. Alle werden gleich. Wie spät mag es sein? Die Macht ist ein laufendes Band, meine Ohren sind Wunden. Es ist so flach hier. Muzak für Leichenschauhäuser und Neubauten. Angenehm summend, hinterlässt keine Spuren. Akkordnarben in meinem Gesicht. Es ist so flach hier. Wie spät mag es sein?“***
>
> *Die genaue Zeit – Einstürzende Neubauten (1983)*

Es dauerte Jahre, bis die Musikindustrie sich davon erholte. Einerseits ging sie juristisch gegen Tauschbörsenbetreiber wie Napster vor, andererseits entwickelte sie mit Plattformen wie Spotify Streamingdienste, über die Nutzer für eine monatliche Pauschale auf nahezu alle jemals produzierten Songs zugreifen können.

Bei solchen monatlichen Abonnements ist der Tauschwert nicht mehr an einen einzelnen Song geknüpft, sondern pauschalisiert und an den Zugriff auf die gesamte Musikbibliothek gebunden. Die auf der vorherigen Stufe durch MP3 entmaterialisierte Ware Musik verkommt so zur Datenressource.

Der Mehrwert fließt fast komplett an die Plattformbetreiber. Egal, wie viel Arbeit sie in einen Song stecken, die Musiker erhalten nur noch einen winzigen Bruchteil pro Stream. Die Plattformen sitzen bei der Aushandlung der Tantiemen mit den Musikern am längeren Hebel. Laut einer im Auftrag der CISAC (International Confederation of Societies of Authors and Composers) Ende 2024 veröffentlichten Studie der französischen Consulting-Firma PMP Strategy schütteten die Streamingdienste 2023 beschämende 8,2 Prozent ihrer Einnahmen an Musiker aus [2].

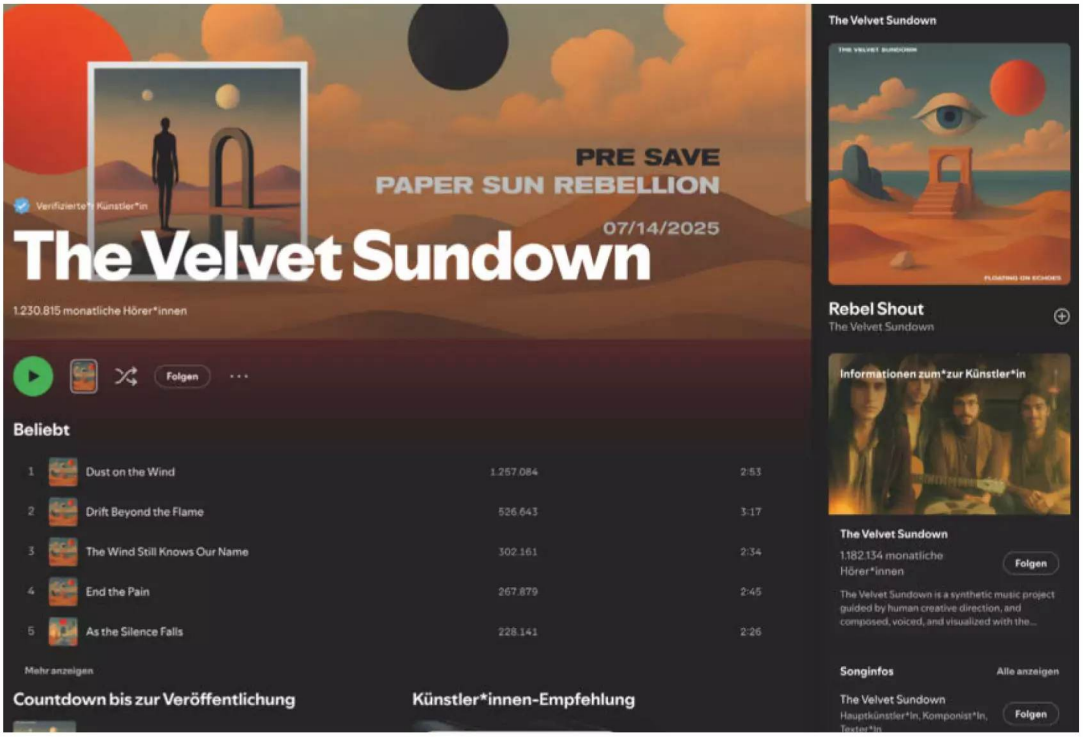
Doch Spotify & Co. sind noch immer darauf angewiesen, tagtäglich mit frischen neuen Songs versorgt zu werden; sonst könnten sie keinen Mehrwert von den Musikern abschöpfen. Bei einer Vollautomation durch KI würde hingegen eine Plattformrente an die Stelle des Mehrwerts treten.

KI-Schwemme der 20er

Laut Geschäftsbericht wurden 2024 jeden Tag 100.000 Songs, also etwa alle 0,8 Sekunden ein neuer Track, auf Spotify hochgeladen. Mit dem Aufkommen von KI-Tools, die neue Songs mit wenigen Klicks produzieren, dürfte diese Zahl in diesem Jahr weiter ansteigen. Erste Vorboten dieser neuen Ära sind KI-Bands wie „The Velvet Sundown“, die auf Spotify dank viraler Social-Media-Kampagnen und einiger Feuilletonberichte innerhalb weniger Wochen über eine Million Follower sammeln konnten.

Alle Songs, Fotos und Plattencover der Band stammen aus KI-Generatoren. Seit ihrem Debüt Anfang Juni 2025 erschien alle zwei Wochen ein neues Album mit rund einem Dutzend weiterer KI-Songs.

Die Macher von „The Velvet Sundown“ erzeugen aber keine neue Musik, sondern in den von ihnen



Die Macher von „The Velvet Sundown“ produzieren mithilfe von KI-Generatoren alle zwei Wochen ein neues Album, wofür echte Musiker ein Jahr oder länger brauchen. Dank der niedrigeren Produktionskosten erzielen sie aus den ausgeschütteten Tantiemen der Streamingplattformen einen höheren Profit, ohne jedoch neue Musik mit Mehrwert zu schaffen.

genutzten Generatoren ist die menschliche Arbeit von Bands wie Led Zepplin oder Fleetwood Mac aus den 70er-Jahren kondensiert, mit denen die KI offenbar trainiert wurde.

Da die Macher den KI-Diensten nur geringe Gebühren zahlen müssen und die Ergebnisse in kurzer Zeit überarbeiten können, erzielen sie durch die Tantiemen, die ihnen die Streamingplattformen auszahlen, einen wesentlich höheren Profit als Produzenten und Verlage, die echte Musiker ein Jahr oder länger an einem neuen Album arbeiten lassen und deutlich höhere Produktionskosten bezahlen.

Die im Vergleich schlechtere Klangqualität der KI-Musik schmälert für die meisten Zuhörer den Gebrauchswert kaum, da sie die Unterschiede auf ihren Bluetooth-Lautprechern oder billigen In-Ears nicht wahrnehmen.

Doch egal ob handgemacht oder KI-generiert, jeder Song bekommt von den Streamingdiensten den gleichen Betrag pro Klick ausgeschüttet. Die handwerkliche, technische oder musikalische Qualität eines Songs spielt keine Rolle. Es geht allein um die Klickzahlen.

Um diese zu erhöhen, müssen die Songs einen möglichst breiten Geschmack treffen und auf beliebten Abspiellisten auftauchen. Die Plattformen generieren solche Listen nicht nur für bestimmte Musikgenres und Epochen, sondern auch für verschiedene Einsatzzwecke im Alltag. Es gibt Abspiellisten zum Training, Joggen, Lernen, Entspannen, Aufwachen, Einschlafen und so weiter – eine moderne Form von Muzak. Laut der PMP-Studie bestehen 41 der 100 meistabonnierten Abspiellisten auf Spotify aus solchen auf Alltagssituationen und Stimmungen zurechtgeschnittenen Muzak-Songs [2].

KI-Klone und Tauschwertverlust

Da die Büchse der Pandora geöffnet ist, wird KI-Musik nicht wieder verschwinden. Selbst wenn die GEMA mit ihrer Klage gegen Suno Erfolg hat, wird sie zusammen mit der die Musiker vertretenden Verwertungsgesellschaft GVL allenfalls eine finanzielle Beteiligung der Komponisten, Texter und Musiker herausschlagen (siehe Seite 138).

Besagte PMP-Studie prognostiziert, dass sich die Umsätze mit KI-Musik von rund einer Milliarde US-Dollar im Jahr 2024 auf 16 Milliarden US-Dollar im Jahr 2028 jährlich verdoppeln. Dann wird bereits jeder fünfte gestreamte Song von Spotify & Co. KI-generiert sein. Zehn Milliarden US-Dollar Tantiemen landen dann nicht mehr bei den Musikern und Komponisten, sondern bei den Produzenten mit KI-Generatoren und deren Anbietern. Bei Musikbibliotheken, die etwa in der Produktion eingesetzt werden, sollen dann sogar 60 Prozent der Aufnahmen aus KI-Generatoren stammen.

Damit das Geschäftsmodell der KI-Plattformen jedoch langfristig funktioniert, benötigen sie Input von Musikern und Komponisten, die neue musikalische Ideen einbringen und Songs zum Training der KI-Generatoren einspielen. Deshalb wird künftig eine steigende Zahl von Musikern für solche KI-Anbieter arbeiten und sie mit neuer frischer Musik versorgen.

Da aber diese Künstler keinen Kontakt mehr zu ihrem Publikum haben, können sie von diesem so wenig identifiziert werden wie die Näherinnen ihrer Turnschuhe. Weil sie als Subjekt verschwinden, werden sie komplett austauschbar und verlieren die Kontrolle über die von ihnen geschaffene Musik. Sie können somit auch das Angebot nicht begrenzen.

Auf „The Velvet Sundown“ werden unzählige KI-Klone mit austauschbaren Produzenten folgen. Das unterscheidet sie beispielsweise von Stars wie Madonna oder Eminem, die die volle Kontrolle darüber haben, wie viele Originalsongs es von ihnen gibt. Durch das steigende Überangebot verliert KI-Musik immer weiter an Tauschwert.

Da die aktuellen Streamingdienste bei der Ausschüttung nicht zwischen KI-generierter und menschengemachter Musik unterscheiden, zieht das auch die Einnahmen realer Musiker und Bands mit runter. Laut der PMP-Studie soll der Anteil, den menschliche Künstler von Streamingplattformen ausgezahlt bekommen, bis 2028 von 8,2 auf 6,4 Prozent fallen. Musiker und Komponisten würden in den kommenden drei Jahren also rund ein Viertel ihrer bisherigen Einnahmen aus dem Musikstreaming einbüßen. In der Folge verarmt der Großteil der Musiker weiter – egal, ob er künftig wie Clickworker für KI-Dienste arbeitet oder für seine Musik weniger ausgezahlt bekommt, weil KI-generierte Songs seine Tantiemen kannibalisieren.

Steal This Album!

KI-Musik hat in den vergangenen drei Jahren große Fortschritte gemacht. Das Niveau entspricht derzeit gefälligen Chart-Songs in lauwarmer MP3-Qualität. Sie dürfte aber spätestens in den 2030er-Jahren zumindest bei der Klangqualität High-End-Niveau erreichen. Da KI die Lohnkosten solch aufwendiger Produktionen erheblich senkt, können alle Verlage

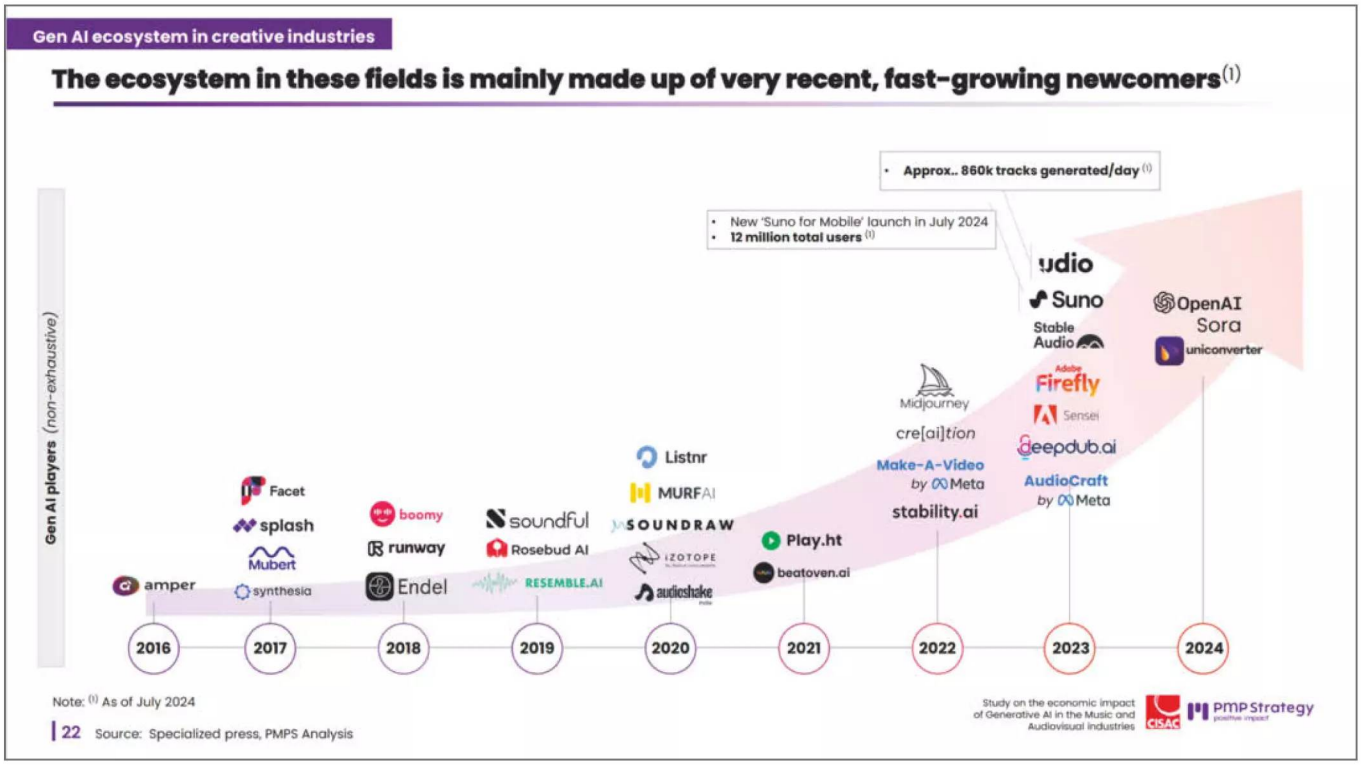
und Künstler, die KI-Generatoren einsetzen, einen deutlich höheren Profit erwirtschaften als diejenigen, die sich ihnen verweigern. Wie heutzutage niemand mehr Wäsche von Hand wäscht, sondern in jedem Haushalt eine Waschmaschine nutzt, könnte auch der Einsatz von KI-Generatoren zur Norm werden.

Das Rennen ist eröffnet: Wer wird dann der größte und beste KI-Dienstleister sein? Die Machtpositionen im Oligopol der Streaminganbieter werden bereits neu ausgekämpft. Aktuelle KI-Start-ups wie Suno und Udio sorgen schon jetzt für diese nächste Phase vor, indem sie in ihren Lizenzbedingungen nahezu unbegrenzte und unwiderrufliche Nutzungsrechte aller mit ihren Generatoren produzierten Songs und aller auf ihre Dienste hochgeladenen Musikmaterialien für sich beanspruchen (Seite 128).

Sollten diese Klauseln rechtlich Bestand haben, dann wird in zehn Jahren kaum noch ein Song industriell produziert und weltweit veröffentlicht, an dem diese KI-Firmen keine umfangreichen und unwiderruflichen Nutzungsrechte haben. Es wäre die größte und umfassendste automatisierte Aneignung musikalischer Werke in der Geschichte. Während der Urheberschutz 70 Jahre nach dem Tod eines menschlichen Komponisten endet, haben die AGBs der maschinellen KI-Dienste kein Ablaufdatum eingebaut.

Rage Against the Machine

Doch wie schon zu Römerzeiten wird es kleine gallische Dörfer geben, die sich der ökonomischen Herrschaft der KI-Dienste über die Musikwelt widersetzen. Auf Entwicklerseite werden vielleicht Open-Source-Projekte mit KI-Tools entstehen, die bald ein ähnliches Niveau wie die kommerziellen Abodienste erreichen. Diese Tools können dann von Musikern kostenlos eingesetzt werden, ohne dass sie die Kontrolle über ihre Musik verlieren.



(Bild: PMP Strategy)

Jährlich poppen immer mehr KI-Musikdienste auf. Neben Start-ups wie Suno und Udio mischen auch große IT-Konzerne wie Meta und OpenAI mit.

Zwar wird man anfangs noch große Server benötigen, doch irgendwann laufen derartige Tools auch auf Smartphones – oder sie werden gleich in In-Ear-Kopfhörer und Bluetooth-Lautsprecher eingebaut. Anwender könnten dann Musik, die sie gerade zur Untermalung ihrer Alltagssituation benötigen, auf Knopfdruck selbst erzeugen – so wie sie auf ihrer Waschmaschine nur noch einen Knopf drücken, um saubere Wäsche zu erhalten. Sie müssten dafür weder Musiker noch KI-Dienstleister bezahlen.

Muzak würde spätestens dann den letzten Rest ihres Tauschwerts komplett verlieren. Damit würde die Nachfrage nach den KI-Musikdiensten zusammenbrechen. Deren Kapitalgeber würden sich andere Investitionsmöglichkeiten suchen. Die auf KI-Muzak aufbauende Musikindustrie würde zwangsläufig absterben, weil mit dieser Musik schlichtweg kein Geld mehr zu verdienen wäre.

Spätestens dann wäre der Zeitpunkt gekommen, die zu Dinosauriern angewachsenen Konzerne zu vergesellschaften. So wie die Lexikonverlage durch Wikipedia abgelöst wurden, würde die Musik zu Creative Commons werden. Damit wäre der Kapitalismus allerdings nicht am Ende, denn wie man bei Wikipedia beobachten konnte, kamen irgendwann clevere Privatunternehmen, die sich das vergesellschaftete Wissen zum Training ihrer Sprachmodelle aneigneten. Aus dem Brockhaus wurde Wikipedia, wurde ChatGPT. Nach dem Kollaps wird sich auch die Musikindustrie wahrscheinlich neu erfinden.

Live Is Life

Und was wird aus den Musikern? Viele Jobs werden in der Tat wegbrechen, etwa die der namenlosen Studiomusiker, die Auftragsarbeiten oder Trainings-

material für KIs einspielen. Was jedoch bleibt, ist das, was die Musikindustrie seit Erfindung der Schellackplatte vor über hundert Jahren nicht verdrängen konnte: das Live-Konzert. Eine KI kann nicht auftreten und keine menschliche Beziehung zum Publikum aufbauen – auch wenn Stars wie ABBA oder KISS dies über Konzerte mit virtuellen Avataren versuchen, um ihre Rente aufzubessern. Zwar feiert in Japan die künstlich quietschende Hatsune Miku seit fast 20 Jahren Erfolge, jenseits der östlichen Manga-Szene blieben sie jedoch aus.

Denn der Wunsch, vor Publikum aufzutreten, seine Emotionen über Musik auszudrücken und mit anderen Menschen zu kommunizieren – nicht zuletzt die Hoffnung, wie ein Star verehrt zu werden –, all dies wird auch im Zeitalter der KI in hundert Jahren noch Musiker und Zuhörer zusammenbringen. Für Live-Konzerte braucht es denn auch Instrumente und Interpreten, die diese spielen können. Das rettet nicht zuletzt die Zünfte der Instrumentenbauer und Musiklehrer.

Was seit Jahrhunderten gilt, wird auch in Zukunft entscheidend sein. Damit Musiker als Subjekt wahrgenommen werden und überleben, müssen sie Aufmerksamkeit erregen und eine direkte Beziehung zu ihrem Publikum aufbauen. Der Weg dorthin führt vielleicht über Wohnzimmerkonzerte, Auftritte in kleinen Clubs und endet für ganz wenige in großen Sälen und Stadien. Dabei ist es umso wichtiger, einen eigenen, am besten einzigartigen Stil zu entwickeln, denn nur so behalten Musiker die Kontrolle über das Angebot und den Tauschwert ihrer eigenen Musik.

Neubauten statt KI

Einen solchen eigenen Stil kann man jedoch nicht mit KI-Generatoren entwickeln, weil diese lediglich nach dem Wahrscheinlichkeitsprinzip bekannte Vorlagen und Stile imitieren, die besonders weitverbreitet, populär und austauschbar sind. KI-Musik wird zudem niemals knapp, sondern im Überfluss vorhanden sein, weshalb ihr Tauschwert gen null tendiert.

Musiker, die in der Flut an KI-Musik nicht untergehen wollen, sollten deshalb ihre Musik nicht mit KI-Generatoren produzieren. Als Beispiel dient etwa die Band Einstürzende Neubauten, die sich vor über 40 Jahren mit Presslufthämmern erfolgreich gegen die damalige Muzak-Schwemme wehrte.

Aus ihrem ureigensten Interesse täten auch die Musik-Streamingdienste gut daran, ihre Ausschüttungen für KI-Musik drastisch zu reduzieren und jene für von Menschen eingespielte Musik zu erhöhen. Andernfalls befeuern sie nicht nur den Qualitätsverfall, sondern untergraben auch ihre eigene Geschäftsgrundlage. Aber vielleicht wäre das für die Musik ja auch eine Befreiung … (hag) c't

Literatur

[1] Florian Butollo, Sabine Nuss, Marx und die Roboter, Vernetzte Produktion, Künstliche Intelligenz und lebendige Arbeit, Dietz Berlin 2023

[2] PMP Strategy, Study on the economic impact of Generative AI in the Music and Audiovisual industries: heise.de/s/AXE4K

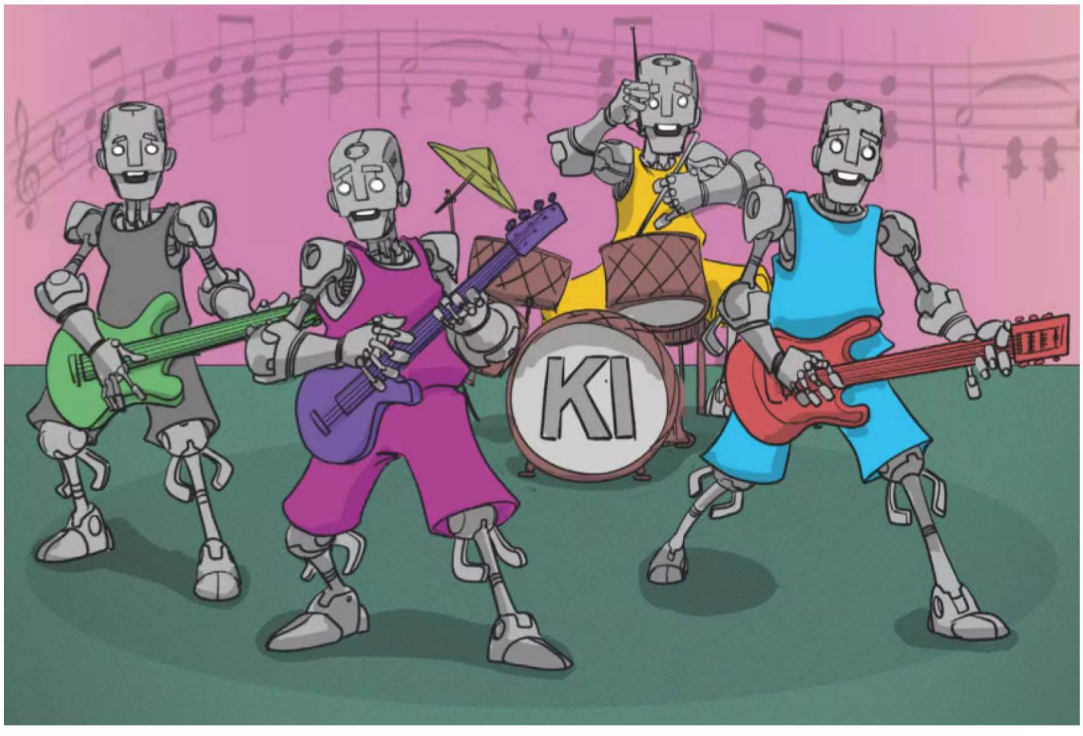


Bild: Thorsten Hübner

Vier KI-Generatoren für Musik im Test

Ein kurzer Prompt – fertig ist der Song. KI-Generatoren füllen zunehmend die Playlisten und machen echten Musikern und Komponisten Konkurrenz. Wir testen nicht nur die Hitqualitäten, sondern decken auch Fußangeln in den AGBs auf.

Von **Kai Schwirzke**

Für unseren Test haben wir vier Musikgeneratoren ausgewählt, die im Netz große Aufmerksamkeit erregen. Allen voran Suno und Udio, die Pop- und Rock-Songs mit Gesang auf kurze Prompts generieren. Boomy konzentriert sich auf elektronische Instrumentalstücke und hat bereits einen Upload-Dienst für Spotify und Apple Music integriert. Demgegenüber spezialisiert sich Aiva auf orchestrale Klänge für Soundtracks. Es berechnet die passenden MIDI-Dateien, sodass man in die generierten Stücke Note für Note gezielt eingreifen kann.

Im Test prüfen wir, inwieweit die Songs der Dienste mit menschengemachter Musik konkurrieren

können. Dabei berücksichtigen wir Melodiösität, den Aufbau der Song-Strukturen, Eingriffsmöglichkeiten für die Anwender sowie die Klangqualität. Weil die KI-Erzeugung von Musik besondere lizenzrechtliche Probleme und Plagiatsklagen nach sich ziehen kann, gehen wir auf die aktuelle rechtliche Situation in einem gesonderten Artikel ab Seite 138 ein. Welche Auswirkungen die KI-Generatoren auf die gesamte Musikindustrie haben und was Musiker gezielt tun können, um nicht durch Algorithmen ersetzt zu werden, beleuchten wir im Artikel auf Seite 120.

Die Preise für KI-Musikgeneratoren reichen von kostenlosen Probe-Abos über Standardlizenzen für rund 12 Euro bis hin zu 36 Euro pro Monat für Pro-Konten, die mehr Funktionen freischalten und Downloads in höherer Stückzahl und Qualität erlauben. Boomy, Suno und Udio liegen preislich gleichauf. Aiva verlangt etwa 50 Prozent mehr, weil es zusätzlich MIDI-Spuren berechnet und exklusive Nutzungsrechte gewährt. Alle Dienste lassen sich im Browser oder über mobile Apps (Android/iOS) bedienen. Die Apps vereinfachen lediglich die Anmeldung, ansonsten bieten sie keine nennenswerten Vorteile, zumal die Songs eh auf fernen Servern errechnet werden. Im Browser hemmten gelegentlich Browsererweiterungen, welche die Tonausgabe beeinflussen, die Abspielfunktionen der Dienste. Man schaltet sie am besten ab. Die Untersuchungen für diesen Test fanden Anfang Juli 2025 statt.

Sag, was du willst

Suno und Udio arbeiten nach einem Text-to-Music-Prinzip: Man schreibt hin, was die KI produzieren soll. Auch Instrumentierung und Stilistik lassen sich im Prompt angeben. Ein Beispiel: „Dark electronic pop song with Metal influences, heavy guitars and atmospheric synth structures." Wer will, kann die Anweisungen auch auf Deutsch formulieren.

Um effektiv zu arbeiten, sollte man wissen, welche Tags und Metatags die KI versteht. Suno beispielsweise erlaubt es, Songstrukturen in eckigen Klammern vorzugeben, etwa [Intro], [Verse], [Chorus]. Innerhalb dieser Struktur sind weitere Differenzierungen möglich. So lässt sich die Instrumentierung variieren oder von männlicher zu weiblicher Stimme wechseln.

Das klingt einfacher, als es in der Praxis ist, zumindest wenn es komplexer werden soll. So baten wir Suno, ein orchestrales Stück über eine herbstliche Abendstimmung mit am Ende einsetzender Frauenstimme zu generieren. Die sollte nicht auf Text, sondern nur bestimmte Vokale oder Konsonanten summen. Kurzum: Es hat nicht geklappt, obwohl es das Toolset prinzipiell hergibt. Bei diesem Beispiel brauchte es auch mehrere Anläufe, Sunos Hang zu schwülstigen Heavy-Gitarren zu bändigen. Diese Schwierigkeiten lassen sich ohne Weiteres auch auf Udio übertragen.

Klick your Song

Boomy und Aiva verfolgen einen anderen Ansatz, der eher einem Musikbaukasten ähnelt. Hier wählt man zunächst aus den vorgegebenen Stilrichtungen die Gewünschte und verfeinert sie dann anhand eines Parametersets. Das kann bei Boomy soweit gehen, dass Sie einzelne Instrumente innerhalb eines Drumsets spezifizieren dürfen, ohne dass das Ganze allzu unübersichtlich wird.

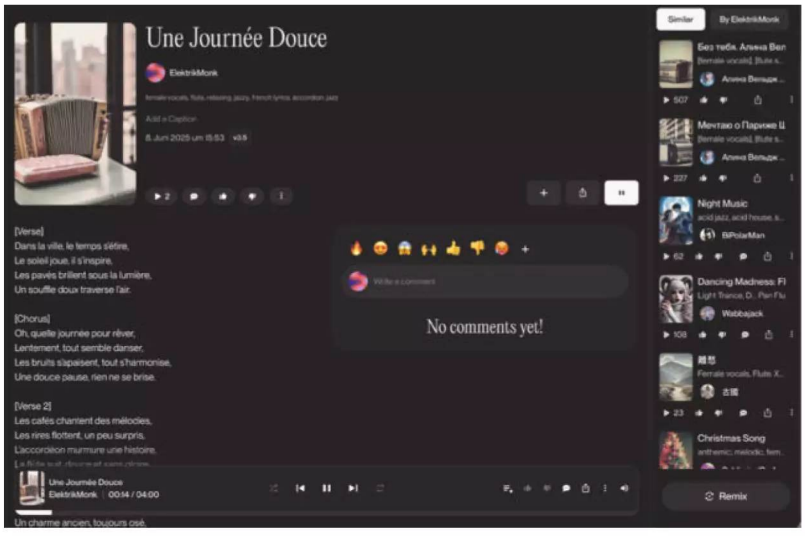
Das schrittweise Zusammenklicken einzelner Aspekte einer Komposition in Aiva dürfte unerfahrenen Anwendern schwerer fallen als musikalisch vorgebildeten. Anwender müssen sich durch eine unüberschaubare Vielzahl an Vorlagen wühlen, bis die passende gefunden ist. Neben Stilen des Herstellers findet man auch Stilvorgaben von anderen Anwendern. Gut gefällt uns, dass sich die Komposition bereits mit Bordmitteln in einem DAW-ähnlichen Editor anpassen und korrigieren lässt. Das klappt auch ohne Bezahl-Abo.

KI singt Deutsch

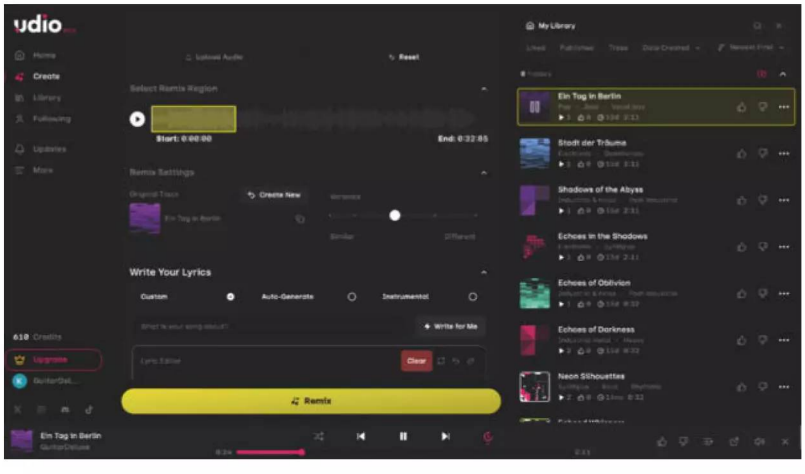
Suno und Udio können im Unterschied zu Aiva und Boomy auch singen, Liedtext inklusive. Dabei gibt man das Songthema einfach im Prompt zusammen mit der gewünschten Stimmlage eines Sängers oder einer Sängerin an. Das sieht dann beispielsweise so aus: `Relaxing song about a chilling day in the city. Jazzy vibes, german vocals`. Wer mag, kann auch seinen eigenen Liedtext hochladen und singen lassen.

Die Dienste trällern nicht nur Englisch, auch Deutsch, Französisch oder Spanisch stehen unter anderem zur Wahl. Das klingt verlockend, sollte aber nur von der Zielsprache halbwegs mächtiger Personen genutzt werden. Denn so inspirierend die Ergebnisse ausfallen können, die KIs greifen häufig daneben. So textet Suno in einem Song über einen chilligen Tag in der Großstadt unfreiwillig komisch:

Die Menschen flanieren, so sorglos frei, ein friedlicher Tanz im Alltagsbrei.



Innerhalb eines Dienstes dürfen User ihre Kreationen untereinander vorstellen, hier bei Suno.



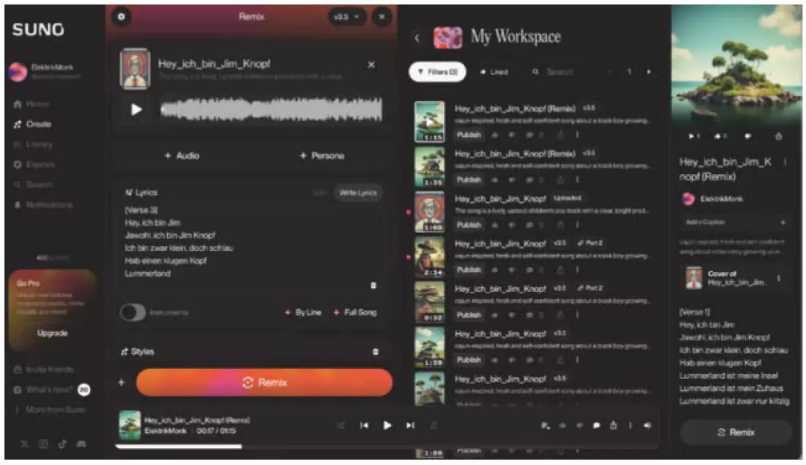
Mit der Remix-Funktion in Udio lässt sich auch die Struktur des Songs nachträglich verändern.

Durch das Versmaß nicht erklärbare Grammatikfehler muss man öfter händisch korrigieren. Bei: „Wo jeder Schritt ein Rhythmus hat“ lässt sich beispielsweise der fehlerhafte Nominativ „ein“ problemlos durch das „nen“ im Akkusativ ersetzen. Womit keinesfalls gesagt sein soll, dass von Menschenhand verfasste Liedtexte stets das Nonplusultra sprachlicher Präzision widerspiegeln, manchmal bestimmt auch die künstlerische Freiheit.

Genaueres Hinhören erfordert ferner die Text-to-voice-Umsetzung der KI. Um beim Großstadtsong zu bleiben: Den „Alltagsbrei“ vermurkste Suno in der dritten Strophe zu etwas Ähnlichem wie „Alter sprich“, nur um etwas später das Wort perfekt auszusprechen.

Struktur, bitte

Ein großes Manko komponierender KI war bislang, ein Musikstück stimmig über die Zeit zu entwickeln, ohne in einförmiges Dudeln zu kommen. Hübsche acht Takte zu schreiben ist das eine, diese Idee in ein gelungenes Musikstück zu verwandeln, etwas ganz anderes. Dazu bedarf es vor allem einer Idee von Form. So besitzen die meisten Popsongs etwa eine Einleitung (Intro), die zur Strophe führt. Dann



Besitzt man die entsprechenden Rechte, kann man sich von Suno eine neue Version eines Songs erstellen lassen.

folgen Refrain, weitere Strophen, ein Zwischenteil et cetera – und zu guter Letzt der Schluss. All dies sollte im Idealfall durch eine musikalische Idee miteinander verbunden sein.

Im Vergleich zu unserem Test von 2023 haben die KIs in dieser Beziehung einiges dazugelernt [1]. Vor allem Suno liefert gut strukturierte Songs, deren Elemente schlüssig ineinandergreifen. Allerdings bereitet der Schluss offenkundig größere Probleme. Stücke zerfasern am Ende gerne. Es klingt dann, als klöppele die KI hilflos antrainiertes „Wissen" zusammen. Abhilfe schaffen zwar Metatags, mit denen sich das Ende beispielsweise als Fade-out gestalten lässt, besonders elegant ist das jedoch nicht.

Udio gelingen ebenso stimmige Strukturen, wiederum mit Problemen am Schluss. Im Vergleich zu Suno wirken Songs in ihrem Aufbau weniger organisch. Rudimentär klingen die formalen Strukturen bei Boomy. Wo Suno und Udio vergleichsweise einfache Tricks wie das Ausdünnen der Instrumentierung beherrschen, spielt Boomy munter volle Pulle. Auch Intros und Outros kann dieser Dienst deutlich schlechter.

Allerdings: Bei Boomy, Suno und Udio kann man einmal erzeugte Songs in deren Editoren nachbessern und etwa weitere Strophen, neue Intros oder Zwischenteile generieren lassen. All das bezahlt man mit sogenannten Credits, einer virtuellen Währung, die die Dienste zur Abrechnung nutzen.

Aiva bekommt ebenfalls einen brauchbaren Liedaufbau hin, zeigt sich dabei aber nicht in dem Maße Pop-orientiert wie die anderen. Das liegt nicht zuletzt daran, dass die Stärken dieses Diensts nicht im Melodischen liegen.

Wie geht das?

Kommerzielle KI-Anbieter lassen sich nur ungern in die Karten schauen. Auf welcher Datenbasis Suno & Co. Songs errechnen, erfährt man selbst auf beharrliches Nachfragen nicht. Ebenso bleibt letztlich unklar, wie es den Diensten gelingt, ansprechende Einzelspuren in dieser stilistischen Vielfalt und Qualität auszuspielen.

Wir sind sicher, dass sie dabei KI-Systeme und konventionelle Algorithmen für das musikalische Regelwerk kombinieren. Die KI benötigt man beispielsweise, um Songtexte zu generieren und diese singen zu lassen. Auch die Analyse und Umsetzung von User-Prompts erledigt die KI.

Der eigentliche Song setzt sich mit hoher Wahrscheinlichkeit aus vorproduzierten Versatzstücken (Patterns) zusammen, die eine KI wiederum nach den Vorgaben des Anwenders zusammenstellt – unterstützt von strikt regelgestützten Rechenroutinen. Letztere sorgen dafür, dass die KI sich nicht allzu weit vom Erwartungshorizont der Hörer entfernt.

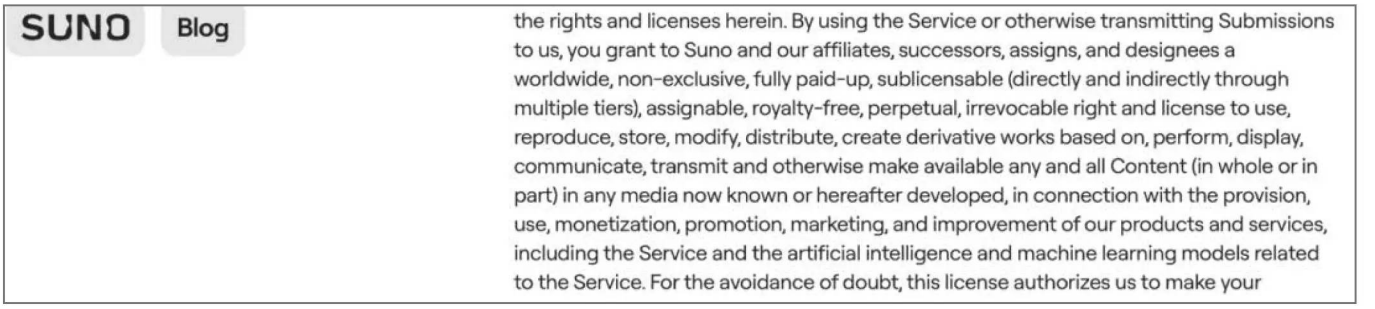
Komponieren nach Vorbildern

Abgesehen von Boomy bieten die Dienste an, Audiodateien als Referenz für neue Kreationen hochzuladen. Bei Suno heißt das Ganze „Remix" und funktioniert im Prinzip auch so: Die KI analysiert den hochgeladenen Song, erkennt den Text und verändert ihn anhand der Anwender-Tags, indem sie beispielsweise Sänger oder die Instrumentierung tauscht. Auch stilistische Variationen sind möglich sowie das Verlängern eines Musikstücks. Das klappte in unseren Tests erstaunlich gut.

Udio geht sogar so weit, tatsächlich ein ganz neues Stück „im Stile von" zu errechnen. Die Ergebnisse hatten allerdings nur entfernte Ähnlichkeit mit den hochgeladenen Vorlagen und überzeugten uns nicht.

In beiden Fällen gilt jedoch, dass der Anwender die Rechte an den hochgeladenen Songs besitzen muss. Wir haben anhand des Gloria-Gaynor-Klassikers „I Will Survive" geprüft, wie genau es die Anbieter damit nehmen. Suno und Udio wiesen nach einiger Rechenzeit unseren Song mit der Anmerkung zurück, dieses Werk gäbe es bereits.

Auch wenn die Beachtung des Urheberrechts grundsätzlich zu begrüßen ist: Dadurch dürften der-



Suno beansprucht weitgehende Nutzungsrechte in seinen AGBs (Terms of Use), die in Deutschland warscheinlich mit dem § 307 BGB kollidieren: "Durch die Nutzung des Dienstes oder die anderweitige Übermittlung von Beiträgen an uns gewähren Sie Suno und unseren verbundenen Unternehmen, Nachfolgern und Beauftragten ein weltweites, nicht exklusives, vollständig bezahltes, unterlizenzierbares (direkt und indirekt über mehrere Ebenen), übertragbares, gebührenfreies, unbefristetes, unwiderrufliches Recht und eine Lizenz zur Nutzung, Vervielfältigung, Speicherung, Änderung, Verbreitung, Erstellung abgeleiteter Werke, Aufführung, Anzeige, zu übertragen und anderweitig verfügbar zu machen, die derzeit bekannt sind oder in Zukunft entwickelt werden, in Verbindung mit der Bereitstellung, Nutzung, Monetarisierung, Werbung, Vermarktung und Verbesserung unserer Produkte und Dienstleistungen, einschließlich des Dienstes und der mit dem Dienst verbundenen Modelle für künstliche Intelligenz und maschinelles Lernen."

artige Funktionen für viele Anwender deutlich an Reiz verlieren. So bleibt nur, eigene oder rechtefreie Werke als Inspirationsquelle heranzuziehen.

Höchst problematisch ist zudem, dass die KI-Anbieter in den AGBs das Recht für sich beanspruchen, die hochgeladenen Songs zum eigenen Training und für beliebige andere Zwecke weiternutzen zu dürfen – bei Suno sogar unwiderruflich. Künstler sollten hier aufpassen, ob sie dort tatsächlich eigene Lieder, Demos oder Fragmente ihrer Musik hochladen und damit zum KI-Training freigeben wollen.

Aiva zeigte sich an dieser Stelle weniger kritisch und akzeptierte das offenkundig urheberrechtlich geschützte Werk. Allerdings hatten die Resultate derart wenig mit dem Ursprungslied zu tun, dass man sich die Mühe hätte sparen können. Das mag vor allem daran liegen, dass Aiva MIDI-basiert arbeitet und auch entsprechende Dateien als Einfluss akzeptiert. So kann man von der KI neue Varianten berechnen lassen, wenn man mit einer eigenen Idee nicht weiterkommt.

Einzelspuren bearbeiten

Ein großes Manko vieler KI-generierter Songs, die nicht wie die von Aiva auf MIDI-Daten beruhen, war lange Zeit, dass sie nur als fertiger Mix, nicht aber als Einzelspuren der Instrumentierung zur Verfügung standen. Stem Separation nennt sich diese Funktion. Genau das benötigt man allerdings, möchte man die KI-Kreation umstandslos in einer DAW wie Cubase, Logic oder Live weiterentwickeln. Schlagzeug, Bass oder Gesang kann man eben nur dann austauschen oder bearbeiten, wenn sie als unabhängige Tracks vorliegen.

Zwar lässt sich der Missstand durch weitere KI-gestützte Software wie Moises kostenpflichtig beheben – auch einige Audioeditoren (Acoustica) und DAWs (Logic) bieten derartige Funktionen an. Die Verfahren sind jedoch verlustbehaftet und an den Einzelspuren kleben digitale Artefakte.

In unserem Testfeld bieten alle Kandidaten mit Ausnahme von Boomy den Download von Stems – wenn man kostenpflichtige Abos abschließt. Qualitativ sind diese Stems den per Software „entmixten“ Spuren überlegen, da sie weniger Artefakte und Störgeräusche enthalten.

Downloadformate und MIDI

Boomy, Suno und Uido bieten ihre Songs und zum Teil auch Einzelspuren im MP3- und Wave-Format an (siehe Tabelle am Ende). Die höchste Qualität bekommen, kaum erstaunlich, nur zahlende Abonnen-

Sowohl Styles als auch Kompositionen lassen sich in Aiva recht detailliert editieren.

ten. Besonders knickerig stellt sich Boomy in der Free-Version an: Deren Nutzer dürfen nämlich gar nichts herunterladen.

Im Unterschied dazu arbeitet Aiva mit MIDI-Daten. Diese repräsentieren lediglich Tonhöhen und -längen, transportieren aber keine Klanginformationen. Dazu benötigt man separate Klangerzeuger als Hard- oder Software (Synthesizer, Sampler etc.), die eben per MIDI gesteuert werden. Um Aivas Kreationen ohne großen Umstand anhören zu können, bietet der Dienst ein eigenes virtuelles Klangset. Die damit gerenderten Songs lassen sich ebenfalls herunterladen.

Diese Klangqualität reicht zwar aus, um sich einen ersten Eindruck von der Komposition zu verschaffen. Möchte man das Stück jedoch weiterbearbeiten, sollte man professionelle Sample-Bibliotheken beispielsweise für Native Instruments Kontakt oder Steinbergs HALion nutzen.

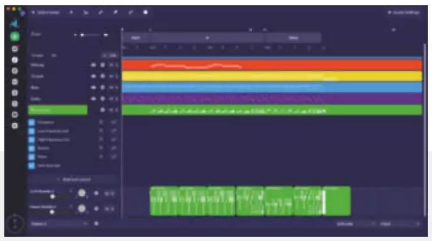
Kurzum: Aiva tönt erst einmal wenig eindrucksvoll. Dafür hat der Anwender aber die vollständige Kontrolle über das Werk, da er es MIDI-Note für MIDI-Note in der DAW seiner Wahl bearbeiten und vor allem erweitern kann. Auch auf die eigentliche Instrumentierung kann er mehr Einfluss nehmen: Missfällt die Gitarre im Chorus, lässt sie sich mit einem Mausklick in ein anderes Modell verwandeln. Das geht so bei den anderen drei Diensten nicht.

Klangqualität und Nachbearbeitung

Vor allem Suno und Udio liefern ansprechende Audioqualität und ebensolche Mixe, die sich ohne viel Federlesens beispielsweise als Hintergrund- oder Gebrauchsmusik in Lokalen, Geschäften oder auch für Werbeclips oder Videoproduktionen eignen. Um audiophile Highend-Produktionen handelt es sich allerdings nicht. Sie klingen mit ihrer dürftigen räumlichen Abbildung und den teilweise auftretenden Artefakten eher wie MP3s mit 128 Kbit/s – selbst wenn man unkomprimierte Wav-Dateien der KI-Songs herunterlädt. Wer Mittel und Möglichkeiten hat, kann die Songs im (Heim-)Studio besser produzieren.

Das allerdings dauert deutlich länger. Die KI-Mixe setzen Effekte eher sparsam ein. Wer elfenhafte Chormusik mit Streichern bestellt, muss sich nicht wundern, wenn eine ordentliche Portion Hall mitgeliefert wird. Doch auch dann hielt sich die Effektdusche bei unseren Versuchen in Grenzen. Lädt man Einzelspuren herunter, gibt es keine „trockenen“ Signale, zumindest die Track-Effekte sind bereits eingerechnet.

Da Suno und Udio für Einzelspuren nur Audiodaten und keine MIDI-Noten liefern, ist es nicht ganz trivial, einzelne Sounds auszutauschen. Dazu muss



Aiva

Aiva generiert im Unterschied zu den anderen drei Diensten Arrangements aus MIDI-Noten. So kann man leicht für jede Spur und jedes Instrument den Sound austauschen und Änderungen vornehmen – das klappt in Aiva über einen Pianorollen-Editor. Zwar kann man die Stücke auch als Audiodateien herunterladen, deren Qualität ist aber mau, da Aiva für die Instrumente nur Standard-MIDI-Bibliotheken einsetzt. Damit es gut klingt, sollte man die MIDI-Spuren in einer Digital Audio Workstation (DAW) mit besseren Software-Instrumenten verknüpfen.

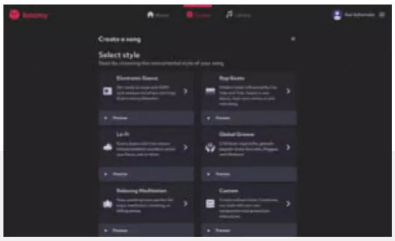
Die Vorbereitungen für einen Song sind aufwendiger als bei der Konkurrenz und setzen eine gewisse musikalische Vorbildung voraus. Aiva stellt für unterschiedliche Stücke und Genres sogenannte Styles zur Verfügung. Die Suche nach einem passenden Stil ist aber mühselig und am Ende generiert die KI nicht immer das, was man im Sinn hatte.

Unserer Ansicht nach eignet sich Aiva besonders gut zum Erstellen orchestraler Texturen, wie sie etwa in der Filmmusik oder auch in Videospielen häufig zum Einsatz kommen. Bei popmusikalischen Stücken klingt die KI hingegen häufig uninspiriert und dudelt belanglos vor sich hin. Aivas komponierte keine griffigen Melodien, das konnte Suno deutlich besser.

Mit einem kostenlosen Benutzerkonto kann man unbegrenzt Musik generieren, darf aber lediglich MIDI- und Audiospuren für drei Songs im Monat herunterladen. Bei einem Pro-Konto für 59 Euro pro Monat überträgt Aiva die vollen Exklusivrechte der Songs an den Nutzer.

- **generiert MIDI-Spuren**
- **gute orchestrale Arrangements**
- **Pro-Plan mit Exklusivrechten**

Preise: kostenlos /
18 oder 59 Euro pro Monat



Boomy

Boomy wendet sich vor allem an Liebhaber instrumentaler Elektronikstücke, die keine weiteren Vorkenntnisse haben. Mehr als zwei bis drei Klicks braucht es nicht, bis ein neuer Song nach ein paar Minuten im Nutzerkonto liegt. Bereits beim ersten Besuch der Webseite nervt der Dienst jedoch mit aufdringlicher Werbung, die wesentliche Teile des Bildschirms verdeckt, sich nur schwer entfernen lässt und bald wieder von Neuem auftaucht.

Die musikalische Auswahl ist dürftig. Boomy bietet lediglich fünf Hauptstilistiken verschiedener Dance-Floor-Richtungen, in denen maximal sieben Unterkategorien zur Auswahl stehen. Im Unterschied zu Suno und Udio generiert Boomy keine Texte und keinen Gesang. Poppige oder gar rockigere Stile sucht man vergebens.

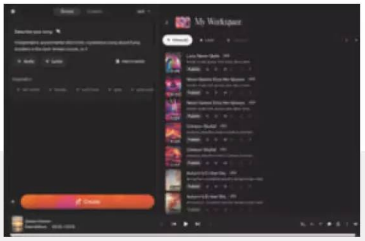
Am ehesten konnten uns ruhigere Ambient-Kompositionen überzeugen. Allzu oft rumpelte die KI jedoch unstrukturiert durchs musikalische Rübenbeet. Das ließ sich auch mit der Schaltfläche „Custom“ kaum verbessern, die Anpassungen des Tempos oder der Instrumentierung erlaubt. Bezahlkunden dürfen die Songs in einem einfachen Editor bearbeiten.

Im Unterschied zu den anderen Anbietern kann man bei Boomy keine Einzelspuren als Stems herunterladen. Trotz unkomprimierter WAV-Dateien bleibt die Klangqualität hinter der von Suno und Udio zurück.

Nutzer können ihre fertigen Boomy-Songs in Playlisten auf der Homepage des Anbieters präsentieren. Doch die Ergebnisse sind dürftig. Nichts klang dort auch nur annähernd interessant.

- **einfache Songerstellung**
- **schlechte Ergebnisse**
- **problematische AGB-Klauseln**

Preise: kostenlos /
12 oder 36 US-Dollar pro Monat



Suno

Geht es darum, mit verblüffend geringem Aufwand ansprechende Songideen oder komplette Titel zu erstellen, hat Suno die Nase vorn. Der Dienst generiert aus einem Textprompt ein komplettes Musikstück, inklusive Text und Gesang.

Nichtmusiker kommen durch die intuitive Text-to-Music-Funktion ebenso rasch zu guten Ergebnissen wie ambitionierte Produzenten, die durch geduldige Promptnutzung Tempo, Tonart, Songstruktur und grundlegende Instrumentierung angeben können.

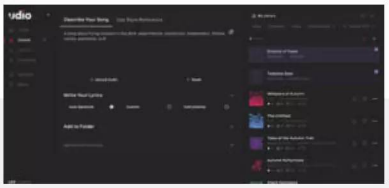
Suno überzeugte uns durch seine große stilistische Bandbreite, sei es bei einem relaxten Großstadtsong in deutscher Sprache oder einem charmanten französischen Straßenchanson mit Akkordeon, Querflöte und Gitarre. Zwar gelingt auch Suno nicht alles gleich gut, die Hände über dem Kopf zusammenschlagen mussten wir allerdings nie.

Die Abrechnung erfolgt wie bei Udio über sogenannte Credits, von denen man auch im kostenlosen Abo täglich einige dazu bekommt. Jede neue Berechnung – auch die der Stems – kostet Credits, egal, ob das Ergebnis verwendbar ist oder nicht. Wir benötigten mehrere Anläufe, bis Suno beispielsweise verstanden hatte, dass im Intro oder Outro niemand singen soll.

Wie bei Udio kann man neben den kompletten Songs auch Einzelspuren (Stems) herunterladen. Diese sind allerdings bereits mit allen Effekten versehen. Trockene Intrumentenspuren, die man in einer DAW besser mischen könnte, gibt es nicht. Zu beachten ist, das Suno in den AGBs darauf pocht, alle generierten und hochgeladenen Songs selbst frei verwenden zu dürfen.

- **gute Resultate (Musik und Text)**
- **Einzelspuren und Songeditor**
- **problematische AGB-Klauseln**

Preise: kostenlos / 12 oder 36 US-Dollar pro Monat



Udio

Udio generiert wie Suno Songs und Lyrics über Textprompts und steuert den Gesang bei. Handhabung und Funktionsumfang ähneln Suno, konnten uns aber weniger überzeugen. So stocherte die von Udio ausgewählte Sängerin bei unserem Großstadtsong recht lustlos im Tonvorrat herum und traf auf keine inspirierende Komposition. Auch beim französischen Chanson schwächelte der Dienst. Gut fanden wir allenfalls einige Dark-Industrial-Kreationen, die deutlich härter zur Sache gingen als bei Suno.

Udio bietet einige Parameter, die sich so bei Boomy und Suno nicht finden. So kann man beispielsweise bestimmen, wie stark der Prompt oder die vorgegebenen Lyrics den späteren Song beeinflussen sollen. Wir fanden diese Optionen wenig hilfreich. Sie führten lediglich dazu, allzu schnell die zur Abrechnung genutzten Credits zu verbrauchen.

Udio besitzt ebenfalls eine Remix-Option: Mit ihr lassen sich einzelne Sektionen markieren und anschließend unter Angabe der Varianz neu berechnen. Wer eigene Audiodateien als Style Referenz hochladen möchte, muss zum Pro-Abo greifen.

In puncto Downloadoptionen, Klangqualität und Nutzungsrechte unterschieden sich Udio und Suno kaum voneinander. Udio begrenzt die Länge der generierten Songs allerdings auf zweieinhalb Minuten – bei Aiva sind es bis zu fünf Minuten, bei Suno bis zu acht. Wie auch Boomy und Suno beansprucht Udio umfangreiche Nutzungsrechte an den Songs.

- **einfache Song-Generierung**
- **Einzelspuren und Songeditor**
- **problematische AGB-Klauseln**

Preis: kostenlos / 12 oder 36 Euro pro Monat

man zunächst die Audiospur in MIDI konvertieren. Viele DAWs bieten diese Funktion – bei stark verzerrten Heavy-Gitarren sind die Ergebnisse aber oft dürftig und fehleranfällig, ebenso bei mehrstimmigen Keyboard-Passagen.

Veröffentlichungen

Ein cleverer Schachzug aller Anbieter ist es, für die mit ihrer KI generierten Songs eine interne Hitparade zu führen. Dazu können sogar Nutzer mit kostenlosen Konten ihre Kreationen auf den Webseiten der Dienste veröffentlichen und der Community zur Bewertung vorlegen. Das funktioniert wie in anderen sozialen Medien mit Daumen hoch und runter sowie integrierter Kommentarfunktion. Schneller und billiger können die Anbieter Feedback für das weitere Training ihrer KI-Generatoren kaum generieren.

Boomy bietet darüber hinaus ein Modell zur Monetarisierung an. So fungiert der Dienst auch als Aggregator, der unter gleichlautendem Label beispielsweise auf Spotify und Apple Music zu finden ist. Hat man mit dem Dienst per Mausklick zugestimmt, spielt es freigegebene Werke auf den Streamingplattformen aus und verrechnet die Einnahmen mit dem Anwender, ohne dass man separate Dienste wie Distrokid benötigt. Zahlungen erfolgen ausschließlich auf ein PayPal-Konto. Das lässt sich, schick aufbereitet, auf der Webseite respektive in der App nachvollziehen.

All das kommuniziert Boomy arg versteckt. Vor allem lässt sich auf der Homepage nicht nachvollziehen, nach welchem Schlüssel die – eher spärliche – Gewinnausschüttung erfolgt. Schlecht ist zudem, dass Boomy sämtliche Rechte an den generierten Songs und Einzelspuren behält.

Ebenso beanspruchen die ebenfalls in den USA ansässigen Anbieter von Suno und Udio in ihren AGBs quasi beliebige Nutzungs- und Verwertungsrechte an den generierten Songs. Die Klauseln sind überaus intransparent, ohne klare Zweckbindung und schließen etwa bei Suno einen Widerruf aus. Derartige Vertragsksauseln verstoßen in Deutschland warscheinlich unter anderem gegen § 307 BGB, weil sie die Anwender unangemessen benachteiligen. Im Streitfall müsste man sich jedoch mit den Boomy, Suno und Udio vor US-Gerichten auseinandersetzen.

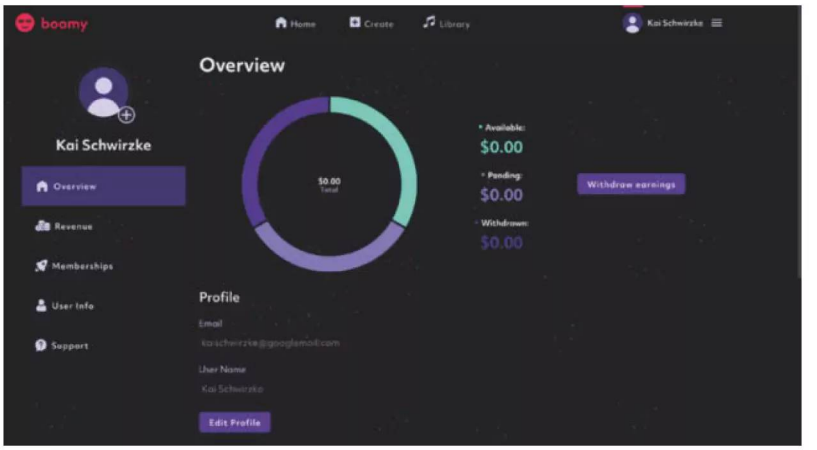
Im Unterschied dazu tritt die in Luxemburg beheimatete Firma Aiva Technologies hingegen Abonnenten des Pro-Plans die vollständigen Rechte an den generierten Songs ab. Diese Rechte sind exklusiv: Aiva selbst darf die generierten Songs nicht weiterverwenden.

Kostenüberblick

Alle Dienste gestatteten es, die Fähigkeiten ihrer KI-Komponisten ohne Abonnement auszuprobieren; dabei zeigt sich vor allem Suno großzügig. Hier darf man seine Kreationen nicht nur als MP3 herunterladen, sondern erhält täglich 50 Credits für bis zu zehn Dreieinhalb-Minuten-Songs. Während unserer Testphase ließen sich diese Credits bis zu einem Guthaben von 500 Credits ansammeln.

Allerdings gilt bei Suno wie bei allen anderen Anbietern: Nur mit einem Abo behält man die kommerziellen Rechte. Rechtssicherheit vor möglichen Urheberrechtsverletzungen bietet das aber keinesfalls (siehe Seite 138).

Ferner sollte man nicht auf die Idee verfallen, die mit seinem kostenlosen Konto erstellten Songs entgegen den Nutzungsbedingungen (also ohne Nennung des KI-Anbieters) auf Instagram, TikTok & Co. hochzuladen. Selbst wenn KI-erstellte Werke nicht urheberrechtlich geschützt sind, so gelten auch hier unter Umständen Leistungsschutzrechte. Missbräuchlich genutzte Songs nachzuweisen, ist für die KI-Anbieter mithilfe selbst von Audioprofis kaum detektierbarer Audiowasserzeichen kinderleicht.



Wie Boomy wie viele Tantiemen an wen auszahlt, haben wir nicht verstanden – und dementsprechend auch nichts verdient.

Anbieter für KI-Musik

Name	Aiva	Boomy	Suno	Udio
Webseite	https://www.aiva.ai	https://boomy.com	https://suno.com	https://uido.com
Hersteller / Land	Aiva Technologies / Lux.	Boomy Corporation / USA	Suno Inc. / USA	Uncharted Labs / USA
Web / App	✓ / Android/i(Pad)OS	✓ / Android/i(Pad)OS	✓ / Android/i(Pad)OS	✓ / Android/i(Pad)OS
DAW-ähnlicher Songeditor	✓	✓	✓	✓
Stem-Download	✓	–	12 Tracks (Pro/Premier)	✓
automatische Lyrics	–	–	✓	✓
KI-Stimme	–	–	✓	✓
Audio-Upload	✓	–	Free: 1 Min. / Standard: 8 Min.	✓ (ab Pro)
Audioformate	MP3/MIDI Pro: zusätzlich HiRes-Wav	Wav	MP3, Wav (nur Pro, Premier)	MP3, Wav (beides nur ab Pro)
Credits pro Monat	–	–	50 (pro Tag), 2500, 10000	10 (pro Tag), 1200, 4800
Downloads pro Monat (free / normal / pro)	3 / 15 / 300	– / 25 / 250	unbegrenzt	unbegrenzt
kommerzielle Nutzung	ab Standard-Plan	ab Standard-Plan	ab Standard-Plan	ab Standard-Plan
exklusive Nutzung	ab Pro-Plan	–	–	–
kostenloser Probeaccount	✓	✓	✓	✓
Preise (im Monat inkl. Mwst.)	Standard: 18 € Pro: 59 €	Creator: 12 US-$ Pro: 36 US-$	Pro: 12 US-$ Premier: 36 US-$	Standard: 12 € Pro: 36 €
Bewertung				
Bedienung	○	○	⊕⊕	⊕⊕
stilistische Vielfalt	⊕	○	⊕⊕	⊕⊕
Songqualität	○	○	⊕⊕	⊕
Klangqualität	⊖[1]	○	⊕	⊕

[1] bezieht sich auf die MIDI-Instrumente des Anbieters

⊕⊕ sehr gut ⊕ gut ○ zufriedenstellend ⊖ schlecht ⊖⊖ sehr schlecht ✓ vorhanden – nicht vorhanden k. A. keine Angabe

Fazit

Künstliche Intelligenz hat sich zu einem ernst zu nehmenden Kompositionswerkzeug entwickelt. Prominente Musiker wie etwa Björn Ulvaeus (Abba) nutzen längst das kreative Potenzial der musikalischen Algorithmen. Sorgfältig durch Könner kuratiert, liefern vor allem Suno und Aiva gute Grundlagen für gelungene Songs.

Hier kommt jedoch das Kleingedruckte der AGBs ins Spiel: Aiva tritt den Nutzern des Pro-Plans die exklusiven Rechte an den generierten Songs ab, sämtliche US-Anbieter tun dies nicht und beanspruchen umfangreiche Nutzungsrechte. Letzteres macht Boomy, Suno und Udio für Künstler problematisch, selbst wenn sie nur KI-generierte Teile in ihren eigenen Songs verwenden wollen.

Was die KI-Generatoren ausspucken, sind vor allem genretypische Songs, deren Aufbau bekannten Mustern folgt, wie man sie zu Hunderttausenden aus Hitparaden kennt. Experimentelle Avantgarde-Musik, die völlig neue Wege geht, fabrizieren Sie abseits von unbeabsichtigten Fehlern und Glitches aber nicht. Nutzer können aber durchaus experimentieren, was beispielsweise passiert, wenn sie Heavy-Metal-Gitarren mit Schlagergesang kombinieren – solche Crossover-Kreationen machen sehr viel Spaß.

Eine wesentliche Anwendung liegt beim ebenso schnellen wie unkomplizierten Generieren von Gebrauchsmusik, etwa um eine Webseite oder ein Video mit ein paar passenden Klängen zu unterlegen. Auf Spotify findet man bereits erste KI-Bands wie „The Velvet Sundown“, die komplett mit Bandporträts, Beschreibungen und Songs mithilfe von KI generiert wurden und über eine Million Abonnenten haben.

Ähnliches gelingt mithilfe von Suno selbst musikalischen Laien, weil es eine deutlich bessere Qualität abliefert als Udio und Boomy. Letzteres ist aufgrund seiner eingebauten Schnittstelle zum direkten Upload bei Streamingdiensten eine prädestinierte Müllschleuder, um sich von der Tantienmenausschüttung eine Scheibe abzuschneiden. Wohin das führen kann, beleuchtet der Artikel auf Seite 120. (hag) c't

Literatur

[1] Kai Schwirzke, Schreib mir einen Song, Vier kommerzielle Musikautomaten mit künstlicher Intelligenz im Test, c't 6/2023, S. 118

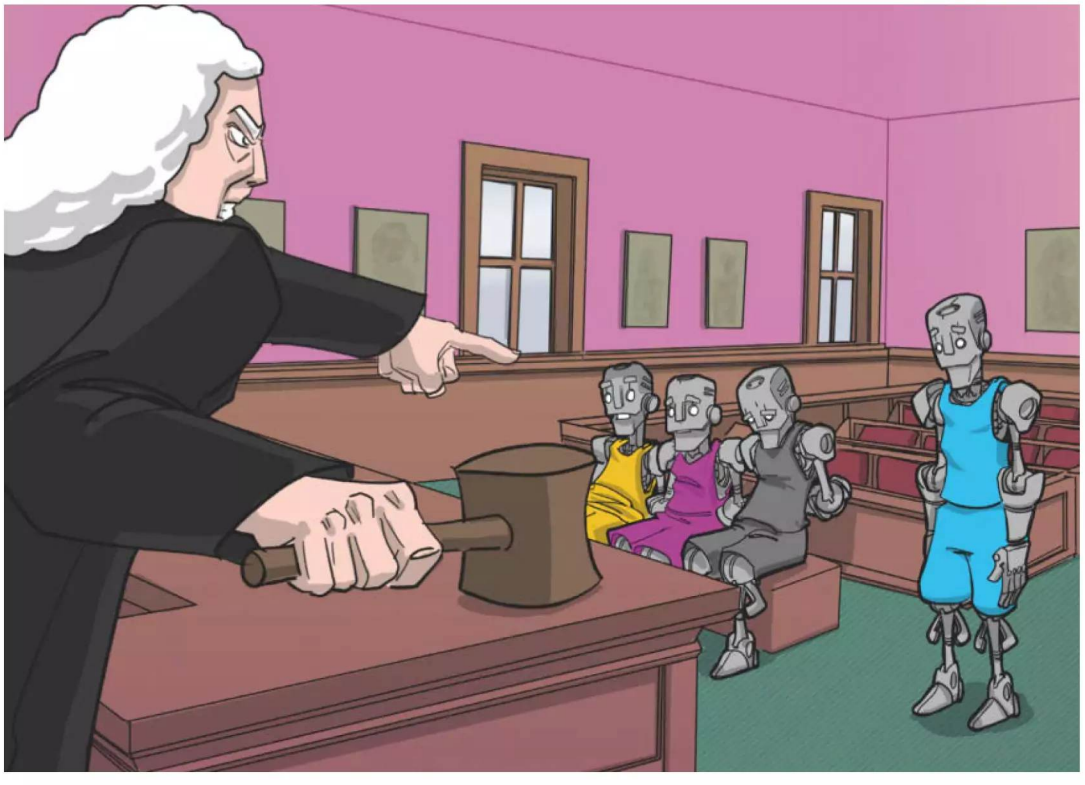


Bild: Thorsten Hübner

Offene Lizenzfragen der KI-Musik

KI-Dienste wie Suno und Udio räumen Anwendern Nutzungsrechte für die generierten Songs ein. Doch was passiert, wenn die Anbieter das gar nicht dürfen, weil sie selbst Rechte von Urhebern verletzen?

Von **Kai Schwirzke**

Eigentlich scheint alles ganz einfach: Man schließt ein Abo mit einem Anbieter von KI-Musik ab und erhält im Gegenzug die uneingeschränkten Veröffentlichungsrechte aller in diesem Zeitraum generierten Musikstücke. Blöd bloß, wenn besagter Anbieter selbst geltendes Recht missachtet.

So ist es nach europäischem Recht verboten, für das Training einer KI urheberrechtlich geschützte Werke zu nutzen, falls deren Schöpfer dieser Nut-

zung ausdrücklich widersprochen haben. Die GEMA hat genau dies stellvertretend für all ihre Mitglieder bereits getan. Der Nachweis eines Verstoßes lässt sich allerdings nur schwer führen. Selbst Musikwissenschaftler können bei Plagiatsvorwürfen unter Komponistenkollegen nicht immer gerichtsfest belegen, dass Song A eine direkte Kopie von Song B ist.

Das liegt unter anderem daran, dass Harmoniefolgen, vor allem weitverbreitete Pop- oder auch Jazzkadenzen, keine ausreichende Schöpfungshöhe besitzen. Die aber ist notwendig, damit ein Werk den Schutz durch das Urheberrecht genießt. Auch sehr einfache Melodien, etwa Volksliedern entlehnt, erreichen die erforderliche Schöpfungshöhe möglicherweise nicht, ebenso manche simple Drum-Grooves.

Schwierige Schöpfungshöhe

Ein Beispiel: Ein deutscher Schlagzeuger versuchte in den Neunzigerjahren vergeblich, den britischen Gitarristen Gary Moore zu verklagen. Es ging um den Hit „Still Got The Blues (for You)“, den der Drummer dem Blueser hinter der Bühne vorgespielt haben wollte.

Das Gericht wies die Klage unabhängig von der Beweislage ab: Bei der Harmoniefolge handle es sich um einen seit Jahrhunderten bekannten Quint/Quart-Fall; die simple Melodie verbinde lediglich sequenzartig die entscheidenden Harmonietöne. Was bedeutet: Es gibt nichts Schützenswertes.

Darüber hinaus stellt es beispielsweise auch keine Verletzung des Urheberrechts dar, „Songs im Stil von“ zu komponieren. Kommt jemand – oder eine KI – auf die Idee, den Sound von Modern Talking mit eigenen Kompositionen aufleben zu lassen, wäre dies völlig legitim – solange nicht bei Dieter Bohlen eins zu eins abgekupfert wird.

Alles unsicher

Für die Abonnenten von KI-Diensten ergeben sich daraus unerfreuliche Unwägbarkeiten. Sie müssen eigenverantwortlich entscheiden, ob es sich beim KI-Song um ein Plagiat handeln könnte.

Bei einem Rechtsstreit oder einer ersten außergerichtlichen Kontaktaufnahme per Abmahnung beziehungsweise Unterlassungserklärung werden Anwälte zunächst denjenigen zur Rechenschaft ziehen, der das vermeintliche Plagiat veröffentlicht hat. Der muss dann nachweisen, dass er das Werk von einer KI hat erstellen lassen und der Dienstanbieter ihm angeblich die Rechte zur Veröffentlichung eingeräumt hat. Bereits dieses Unterfangen dürfte nicht wenige Anwender überfordern.

Doch selbst wenn es ihm gelingen sollte, die KI-Herkunft des streitigen Werks zu beweisen, ist er nicht aus dem Schneider. Handelt es sich dennoch um ein offensichtliches Plagiat, muss der Nutzer mit kostspieligen Schadenersatzansprüchen rechnen. Er hätte sich schließlich, so die vorherrschende Meinung, über die möglichen Risiken einer Veröffentlichung vorher(!) kundig machen müssen.

Anbieterhaftung

So bleibt Nutzern, die verklagt werden, ihrerseits eine Rückgriffsklage gegen den Anbieter anzustrengen. Den in Haftung zu nehmen, kann kompliziert sein. Viele Dienste geben im Impressum keine ladungsfähige Adresse an. Auch Briefkastenfirmen auf den Antillen oder in Alaska dürften hierzulande Beklagten und ihren Anwälten kaum als Ansprechpartner zur Verfügung stehen. Die „Terms of Use“ der drei US-Anbieter Boomy, Suno und Udio scheeren sich nicht um europäisches Urheber- oder deutsches Vertragsrecht und sind zu diesen wahrscheinlich inkompatibel.

Nahezu unmöglich gestaltet sich zudem der Nachweis, dass die KI tatsächlich mit Werken von Dieter Bohlen trainiert wurde, um beim Beispiel zu bleiben. Das wäre der erste Schritt für den Anwender, sich schadlos zu halten. In diesem Fall hätte der Anbieter nämlich erstens das Urheberrecht verletzt, zweitens sich nicht konform zum European AI Act verhalten (Angabe der Trainingsdaten) und es drittens versäumt, auf mögliche Risiken bei der Veröffentlichung eines Songs hinzuweisen.

GEMA und Urheberschaft

Der GEMA, dem deutschen Rechteverwerter für Komponisten und Textdichter, gelang in Experimenten mit der KI Suno Anfang des Jahres genau das: Mit den passenden Textprompts produzierte die KI Songs, die – selbst für Laien nachvollziehbar – bekannten Stücken von Alphaville, Bohlen oder Helene Fischer bis aufs i-Tüpfelchen glichen.

Trotz allem bestreitet Suno bis heute, unerlaubt urheberrechtlich geschütztes Material zum Training einzusetzen. Die Entscheidung des Landgerichts (LG) München über die von der GEMA im Januar 2025 eingereichte Klage gegen Suno Inc. steht noch aus.

Suno hat daraus bereits gelernt: Die auch in Videos dokumentierten Verstöße lassen sich mittlerweile nicht mehr reproduzieren. Das allerdings hilft niemandem weiter.

In der Zwischenzeit müht sich der Rechteverwerter an einer weiteren Baustelle: Momentan sind KI-generierte Werke im Sinne des Urheberrechts nicht schutzfähig. GEMA-Mitglieder dürfen daher KI-Kompositionen nicht unter ihrem Namen anmelden, um in den Genuss von Vergütungen zu kommen, sollte das Werk öffentlich aufgeführt oder im Internet gestreamt werden. Das gilt auch, wenn der KI-Anbieter die Rechte zur Veröffentlichung eingeräumt hat.

Allerdings lässt sich Urheberschaft und somit Schutzwürdigkeit herstellen, sobald ein Textdichter oder Komponist KI nur als Werkzeug einsetzt und den Output kuratiert, also bearbeitet. Konkret kann das beispielsweise durch Erweiterung des Arrangements, Ändern der Melodieführung oder Verfeinerung des sprachlichen Ausdrucks geschehen.

Tantiemen durch Streaming

Das klingt nachvollziehbar; den entsprechenden Nachweis zu führen – respektive das Gegenteil zu beweisen, fällt aber nicht immer leicht. Es überrascht daher kaum, dass findige Zeitgenossen die GEMA mit Anmeldungen ausschließlich KI-generierter Werke fluten. Sinn der Übung: Für Streams erhält der Urheber, sofern er bei der GEMA gemeldet ist, eine Vergütung. Erstellt man also bei den einschlägigen Streaminganbietern (Künstler-)Konten und befüllt diese mit KI-Kompositionen, gibt es für jeden Aufruf Geld von der GEMA.

Zwar scheint der Aufwand angesichts recht magerer 0,3 Cent(!) pro Aufruf abwegig. Bedenkt man jedoch, dass sich über Bots massenhaft Klicks und somit Streams erzeugen lassen und auch seitens der Streamingdienste bei entsprechendem Traffic Geld fließt, sieht die Sache schon anders aus. Pro Monat sind vierstellige Eurobeträge nicht unrealistisch.

Die GEMA wehrt sich gegen diesen Missbrauch durch Prüfung auffälliger Mitgliederkonten: Da nur aufgerufene Streams vergütet werden, weiß die GEMA, wo die in Verdacht stehenden Werke zu finden sind, und lässt diese dann auf möglichen KI-Ursprung untersuchen. Das funktioniert ähnlich wie bei Tools zur Prüfung von Texten, bei denen Algorithmen typische Sprachmuster künstlicher Intelligenzen erkennen.

Bestätigt sich der Verdacht und kann der Betreffende etwa nicht durch Noten oder Audiodateien nachweisen, dass er tatsächlich der Urheber ist, führt dies in schweren Fällen zum Ausschluss aus der GEMA und kann zivil- und strafrechtliche Konsequenzen nach sich ziehen.

KI-Musik auf YouTube

Etwas anders sieht die Situation bei YouTube aus. Der Dienst erlaubt KI-generierte Inhalte und monetarisiert diese bei entsprechenden Abrufzahlen. Allerdings gibt es einige Bedingungen.

Zunächst bedürfen alle „synthetisch erstellten Inhalte“ einer expliziten Kennzeichnung. Das gilt insbesondere für sehr realistisch klingende Imitate real existierender Werke oder Künstler. Beschreibungen wie „Ein neuer Song von Taylor Swift“ sind in so einem Fall nicht gestattet.

Außerdem verlangt YouTube, dass Uploader im Besitz sämtlicher Veröffentlichungsrechte sind. Das umfasst neben dem eigentlichen Song auch das Material, mit dem dieser aufgepeppt wurde, etwa Samples oder Loops.

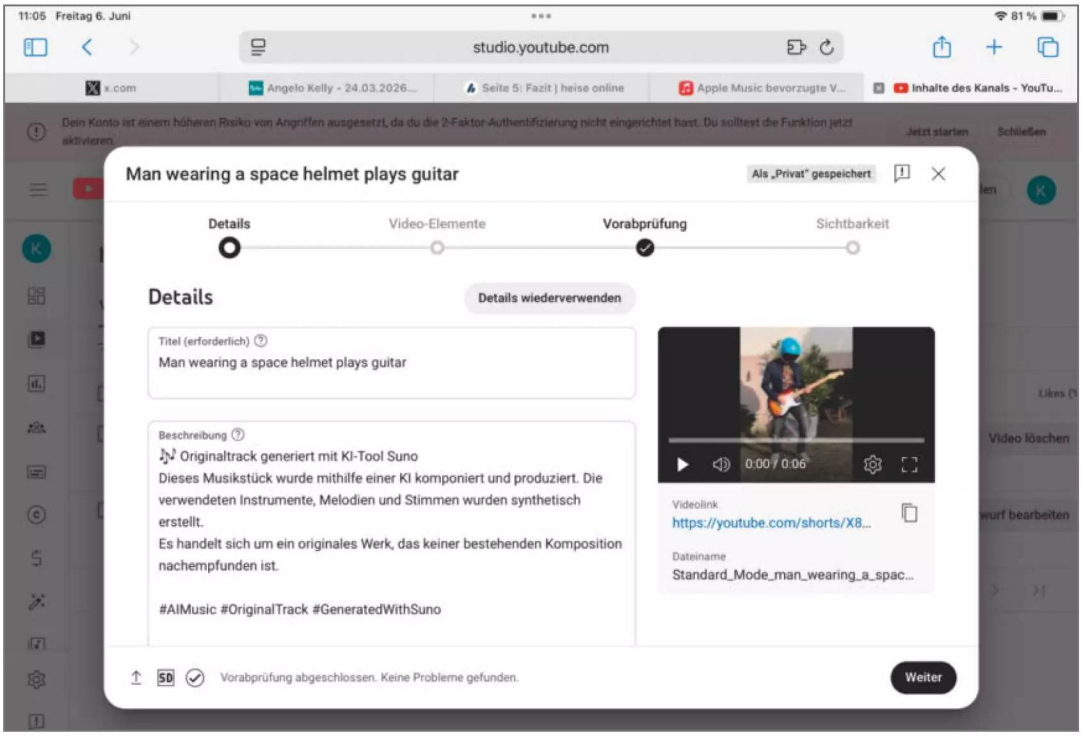
YouTube verbietet ferner den Upload von Songs, die den Verdacht eines Plagiats nahelegen, ob KI-generiert oder nicht. Produziert die KI beispielsweise deutlich erkennbar „Cheri, cheri Lady“, löscht der Dienst das Video oder leitet die Monetarisierung auf den eigentlichen Rechteinhaber um. Passiert dies dreimal, löscht YouTube das Benutzerkonto. Unabhängig davon drohen in solchen Fällen anwaltliche Abmahnungen, die gerade bei gewerblichen YouTubern mit empfindlichen Kosten in fünfstelligem Bereich verbunden sein können.

Spotify und Apple Music

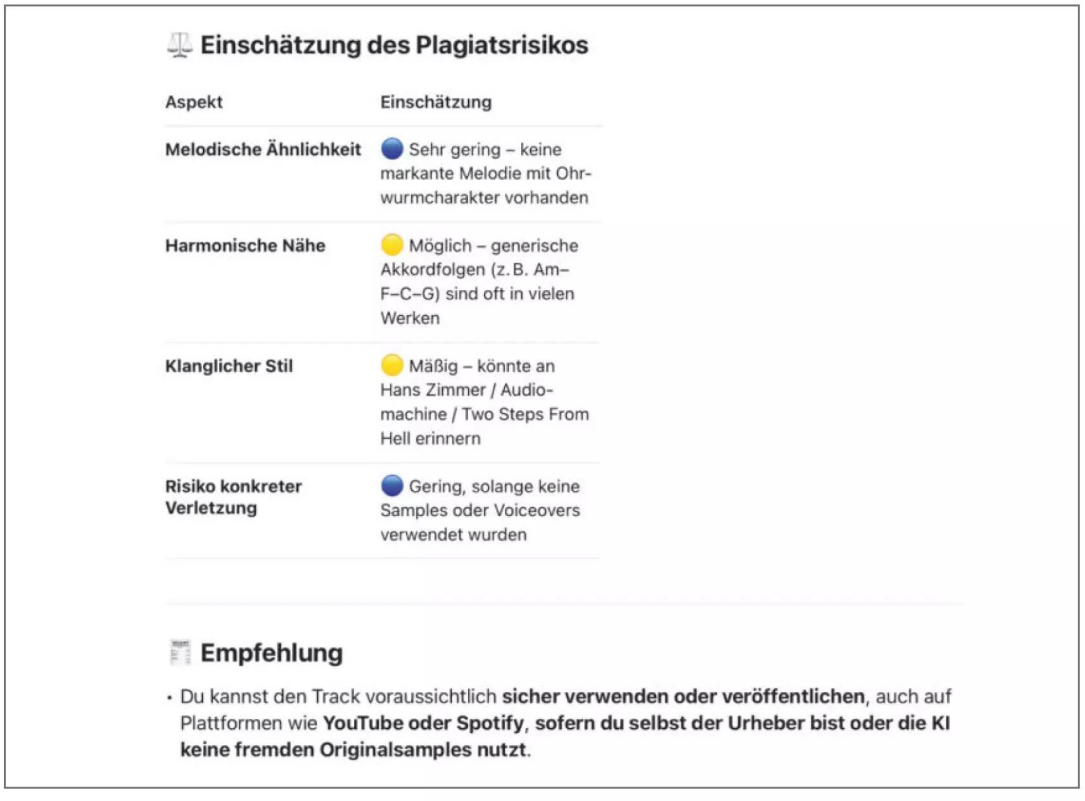
Der schwedische Anbieter Spotify hat grundsätzlich nichts dagegen, wenn „Artists“ von einer KI komponierte Werke hochladen. Allerdings dürfen diese nicht gegen das Urheberrecht verstoßen, und sämtliche Rechte zur Veröffentlichung müssen beim Uploader liegen.

Spotify verbietet in den hauseigenen Richtlinien allerdings Inhalte, die ohne Zustimmung Stil oder Stimme bekannter Musiker imitieren. Ausdrücklich unerwünscht sind außerdem, kaum überraschend, Bots, die Streamingzahlen in die Höhe treiben.

Apple erlaubt ebenfalls KI-Musik, wiederum vorausgesetzt, sämtliche Rechte liegen beim Uploader. Verboten sind auch hier „Klingt wie“-Tracks, ohne dass eine Genehmigung des kopierten Künstlers vorliegt.



Wurden Inhalte per KI erstellt oder verändert, verlangt YouTube in der Beschreibung einen eindeutigen Hinweis.



⚖️ Einschätzung des Plagiatsrisikos

Aspekt	Einschätzung
Melodische Ähnlichkeit	🔵 Sehr gering – keine markante Melodie mit Ohrwurmcharakter vorhanden
Harmonische Nähe	🟡 Möglich – generische Akkordfolgen (z. B. Am–F–C–G) sind oft in vielen Werken
Klanglicher Stil	🟡 Mäßig – könnte an Hans Zimmer / Audiomachine / Two Steps From Hell erinnern
Risiko konkreter Verletzung	🔵 Gering, solange keine Samples oder Voiceovers verwendet wurden

📄 Empfehlung

- Du kannst den Track voraussichtlich **sicher verwenden oder veröffentlichen**, auch auf Plattformen wie **YouTube oder Spotify**, **sofern du selbst der Urheber bist oder die KI keine fremden Originalsamples nutzt**.

Auch ChatGPT analysiert Musikstücke bezüglich ihrer Ähnlichkeit zu anderen Werken. Die Ergebnisse können uns aber nicht überzeugen.

Cupertino setzt die Schwelle für Veröffentlichungen ohnehin ein gutes Stück höher. So verlangt Apple, dass Qualitätsstandards eingehalten werden. Die beziehen sich vor allem auf die Audioqualität der Songs. Rumpelige 64-kBit-MP3s, wie sie viele KI-Probekonten liefern, haben keine Chance auf Veröffentlichung. Außerdem müssen Songs mit einem gewissen Drumherum geliefert werden, etwa Cover-Artworks und aussagekräftigen Metadaten.

Apple unterbindet zudem den direkten Upload von Musik. Hierfür benötigt man einen passenden Vertriebspartner (Aggregator genannt), beispielsweise TuneCore, GoldenDynamic oder DistroKid. Auch hier überprüfen Mensch und Maschine das eingereichte Material auf Rechtsverletzungen.

Sicherheitsgurte

Einer der dicksten Fallstricke bei KI-generierter Musik liegt im allzu unkritischen Wiederkäuen des Trainingsmaterials seitens der Algorithmen. Nicht immer müssen Urheberrechtsverstöße dabei so offenkundig werden, wie von der GEMA vorexerziert.

Da selbst Musikprofis nur einen Bruchteil des urheberrechtlich geschützten Materials kennen können, stellt sich die Frage nach effektiven Schutzmaßnahmen.

Eine Reihe spezialisierter Dienste prüft dafür, wiederum mithilfe von KI, ob ein Musikstück einem anderen verdächtig ähnelt. Wir haben uns einige davon angesehen. Soundcloud nutzt das hausei-

gene Musiio. Normale Anwender erhalten aber lediglich einen sehr eingeschränkten Probezugang. Ziemlich beliebt sind ferner cyanide.ai und audd.io, die sich kostenlos ausprobieren lassen. Sogar ChatGPT und artverwandte Bots bieten an, Audiodateien auf mögliche Übereinstimmungen mit bereits existierenden Werken abzuklopfen.

Dasselbe gilt für alle dieser Tools. So erkannte Audd.io zwar unseren „The Fool On The Hill"-Rip, bei dem wir unter anderem Paul McCartney gegen einen KI-Bowie ausgetauscht hatten. Das offensichtliche Plagiat des Helene-Fischer-Songs „Atemlos" aus den GEMA-Experimenten ließ Audd.io jedoch anstandslos passieren.

Unserem Industrial-Song „Eternal Eclipse" aus Suno mit deutlichen Grindcore-Anleihen attestierte ChatGPT gar eine eher abwegige Nähe zu Hans Zimmer („Time") oder Audiomachine („Guardians At The Gate"). Rechtssicherheit geht anders.

Prüfdienste für Plagiate:
ct.de/wkk3

Fazit

Wer KI-generierte Musik veröffentlichen möchte, muss Sorgfalt walten lassen. Die per AGB zugesicherten Songrechte eines KI-Anbieters enthalten nämlich keinen Freifahrtschein bei eventuellen Ansprüchen fremder Urheber. Solange ungeklärt ist, ob die KI-Dienste ihre Generatoren mit legalen Mitteln trainiert haben, können sich auch deren Anwender nicht darauf verlassen, tatsächlich alle notwendigen Rechte an den generierten Songs zu erhalten. Auf den Kosten für Abmahnungen oder gar Gerichtsverfahren werden sie im Regelfall als Uploader hängen bleiben.

Dienste wie Audd.io helfen zwar, das Risiko zu vermindern, Allheilmittel sind sie aber nicht. Mehr Sicherheit könnte nur eine Rechtsprechung schaffen, die Dienstanbieter stärker in die Haftung einbezieht. Aber auch dann ist es immer noch der Anwender, der den Upload-Button betätigt. (hag) ct

KI-Stimmen: Ruf nach Regulierung

Der Verband Deutscher Sprecher:innen e.V. (VDS) fordert eine umfassende und verbindliche Regulierung für KI-Stimmen in Film, Fernsehen und anderen audiovisuellen Medien. Wir haben die VDS-Vertreterin und Synchronsprecherin Ranja Bonalana gefragt, was an KI-Stimmen so problematisch ist.

Von **Nico Jurran**

c't: Frau Bonalana, in seinem KI-Statement spricht der Verband von der künstlerischen Bedeutung menschlicher Stimmen bei der Vertonung von Filmen, Hörspielen und Spielen. Was ist damit gemeint?

Ranja Bonalana: Bei einer Synchronisation geht es immer um menschliche Kreativität. Und diese beruht auf unseren Emotionen, Vorstellungskraft, auf unseren Lernerfahrungen, Denkprozessen, Reflexionen und so weiter. Unsere Stimmen können all diese Emotionen auf natürliche Weise widerspiegeln. Sie können berühren, verärgern, verängstigen, sind wandelbar und anpassungsfähig. Da wird ja nicht einfach nur auf Deutsch nachgeplappert, was etwa im englischen, italienischen oder französischen Original gesagt wurde. Wir vertonen also nicht nur, wir spielen diese Werke schauspielerisch nach.

» Ich möchte nicht, dass meine Stimme nach meinem Tod für neue Sachen genutzt wird. «

c't: Inwieweit gefährdet KI nun diesen Ansatz?

Bonalana: Der KI fehlt halt jedes Bewusstsein für soziale, emotionale oder gesellschaftliche Faktoren. Sie kann die Emotionen nicht fühlen. Sie weiß ja nicht, was sie sagt. Es ist nur ein Algorithmus, der mit menschlichen Werken gefüttert wurde und bereits Existierendes einfach neu interpretiert und wiedergibt. Und da fehlt es halt komplett an Authentizität und Empathie. Aber wir wollen auf gar keinen Fall die KI verteufeln oder auch um jeden Preis und unreflektiert an Altem festhalten.

c't: Im VDS-Statement heißt es, es ginge um faire Spielregeln und transparente Bedingungen für alle. Wie würden diese in der Praxis aussehen?

Bonalana: Wir wollen nur noch mit ausdrücklichem Einverständnis unsere Stimme für irgendwelche Sachen hergeben. Das heißt, es darf nicht mehr illegal trainiert werden. Wir brauchen definitiv Ausschlussklauseln in unseren Verträgen, da dort momentan einfach alles eingeschlossen ist. Die Verträge sind jetzt bereits absurd lang und werden gefühlt immer länger.

c't: Erleben wir in Zukunft, dass Sprecher und Sprecherinnen klagen müssen, um ihre Stimme zu verteidigen?

Bonalana: Nach aktuellem Stand leider ja. Wir haben auch schon einige Kollegen und Kolleginnen, die klagen beziehungsweise auf dem Weg dahin sind. Das ist aber alles nicht so leicht und sehr langwierig. Gerade, wenn wir über einen Prozess gegen große Auftraggeber reden, ist es wie bei David gegen Goliath.

Deswegen wollen wir eindeutige Regelungen, die das im Vorhinein klären. Ich will als Sprecherin selbst

entscheiden, ob meine Stimme für KI genutzt werden darf – nach dem Motto: „Du musst mich fragen, aber du musst dann dafür auch bezahlen." Wenn die Trainingsdaten Geld kosten, dann wird im Umkehrschluss vielleicht auch gar nicht mehr so viel trainiert. Oder es ist dann auch nicht mehr so leicht, einen Klon zu erstellen, um damit kommerzielle Erfolge zu feiern.

c't: Von KI-Firmen hört man auch oft, dass es doch toll wäre, die Stimmen der Schauspieler zu klonen. Dann würde etwa Keanu Reeves weltweit mit seiner eigenen Stimme sprechen, nur eben halt in den jeweiligen Landessprachen.

Bonalana: Ich glaube, das ist der Traum der Tech-Unternehmen. Als es mit generativer KI losging, kamen sie auf diese Idee und behaupteten, in der Postproduktion direkt 60 Sprachen zu liefern – effizienter, schneller und billiger. Die Produzenten fanden das natürlich total toll, haben aber ganz schnell gemerkt, dass das so nicht funktioniert. Weil sich beispielsweise die amerikanische Sprachmelodie nicht eins zu eins in alle anderen Sprachen übertragen lässt.

Insofern klingt es total albern, wenn ich auf die Originalstimme einfach eine deutsche Stimme draufrechne. Da haben die Produzenten ganz schnell gemerkt: Wir brauchen jetzt doch weiter Sprecher und Sprecherinnen, die das eben nur nicht irgendwie, sondern künstlerisch wertvoll einsprechen – also im Endeffekt nachspielen, da es ja nicht nur um die Stimmfarbe geht, sondern auch um die Stimmführung.

Und es gibt auch Hollywoodstars, die das gar nicht wollen, denn sie sagen: „Seit Jahrzehnten funktioniert es in den anderen Ländern mit meiner dortigen Stimme, warum sollte ich das ändern wollen?" Also: Hugh Grant zum Beispiel sagt eindeutig, dass er seine deutsche Stimme lieber mag als seine Originalstimme. Und auf dem deutschen, französischen und italienischen Markt sind die Leute an die Stimmen gewöhnt. Für die wäre die Originalstimme fremd.

» Wir vertonen nicht nur, wir spielen diese Werke schauspielerisch nach. «

c't: Auf Netflix erscheinen immer wieder – vor allem kleinere – Produktionen, die gar nicht synchronisiert sind.

Bonalana: Weil man sagt, dass dies zu teuer sei. Als Netflix auf den deutschen Markt kam, wollte der Dienst gar nicht synchronisieren. Die Betreiber dachten, sie könnten einfach alles im O-Ton veröffentlichen und die Leute würden es trotzdem konsumieren. Dann haben sie gemerkt, dass das nicht funktioniert und erst einmal wahnsinnig schlechte Synchros auf den Markt gehauen. Doch auch damit sind sie gescheitert. Es gab sogar drei namhafte Produktionen, die neu synchronisiert werden mussten. Und so was kostet alles Geld. Am Ende hat Netflix also gemerkt, dass sie zumindest in Deutschland und in Europa so nicht weiterkommen. Mittlerweile hat Netflix eigene Synchron-Supervisoren.

Bild: privat

Ranja Bonalana ist seit 1984 als Synchronsprecherin tätig und lieh unter anderem Renée Zellweger und Reese Witherspoon ihre Stimme. Daneben ist sie eine gefragte Hörbuchinterpretin. Aktuell kämpft Bonalana mit dem Verband Deutscher Sprecher für den Schutz der Kunst vor KI.

c't: Gibt es für Sie bei KI auch irgendwelche positiven Aspekte?

Bonalana: Das ist schwierig. Wenn alles in einem guten Rahmen reguliert wäre und es ein brauchbares Lizenzmodell gäbe, dann könnte ich mir vorstellen, dass man zum Beispiel Retakes [eine erneute Aufnahme, Anm. d. Red.] mit der KI macht oder sie einsetzt, wenn in der Aufnahme ein Spratzler drin ist oder ein Wort falsch gesprochen wurde. Wir

Der Sprecherverband VDS macht sich in einer Internet-Kamapgne gegen KI-Stimmen stark – unter anderem mit einem Video, das zeigt, wer hinter den deutschen Stimmen bekannter Schauspieler steckt.

bekommen zudem viele große Filme nicht in der finalen Version. Dann fangen wir an zu synchronisieren und erhalten erst ein paar Wochen später das finale Bild. Dann gibt es oft Anpassungen. Das könnte man dann mit der KI machen. Oder wenn jemand im Urlaub ist oder erkrankt.

c't: Nun sterben Synchronsprecher auch ...

Bonalana: Das ist schon eine schwierige Situation. Ich möchte nicht, dass meine Stimme, die ja in vielen Aufnahmen existiert, nach meinem Tod für neue Sachen genutzt wird, bei denen ich keine Hand mehr drauf habe.

c't: Es gibt auch noch die Situation, in der ein Synchronsprecher zwei verschiedene Schauspieler synchronisiert hat, die dann dummerweise in einem Film zusammen auftraten.

Bonalana: Da ist das klassische Paradebeispiel Thomas Danneberg als Arnold Schwarzenegger und Sylvester Stallone. In einem Film hat man sich ja getraut, ihn für beide sprechen zu lassen. Und es ist nicht aufgefallen. Auch, weil wir als Menschen Personen mit Stimmen verknüpften. Wenn ich also die Stimme höre und Sylvester Stallone dabei sehe, dann klingt der für mich wie Stallone – und bei Schwarzenegger ist es genauso. Und es gibt ja auch viele Sprecher, die wirklich extrem unterschiedlich sprechen können.

c't: Im VDS-Statement heißt es, auch Konsumenten müssten für das Thema KI-Stimmen sensibilisiert werden. Wie kann das aussehen?

Bonalana: In erster Linie kann man unsere Petition KunstvorKI unterschreiben, was bislang schon bei über 67.000 Menschen getan haben. Weiterhin kann man sich beschweren, seinen Unmut äußern – also den Verleih, den Verlag oder das Label anschreiben oder dafür Social Media zu nutzen. Kritik mögen die gar nicht.

Und man kann um Stellungnahmen bitten, warum das denn gemacht wird. Im besten Fall Inhalte nicht mehr konsumieren. Dass das Publikum eine unglaubliche Macht hat, zeigt das Beispiel des großen Animeanbieters Crunchyroll. Der hat vor Kurzem ein Statement herausgegeben, dass er in seinen kreativen Prozessen komplett auf KI verzichten wird. Das geschah auch nur aufgrund des Protests der Zuschauer und der Kreativen. (nij) *c't*

Weitere Infos:
ct.de/wt5f

betterCode()

.NET 10.0

Das Online-Event zum neuen LTS-Release

18. November 2025 • Online

Auf dieser Konferenz präsentieren .NET-Experten den fertigen Stand des neuen Release mit drei Jahren Long-Term Support.

Das Programm:

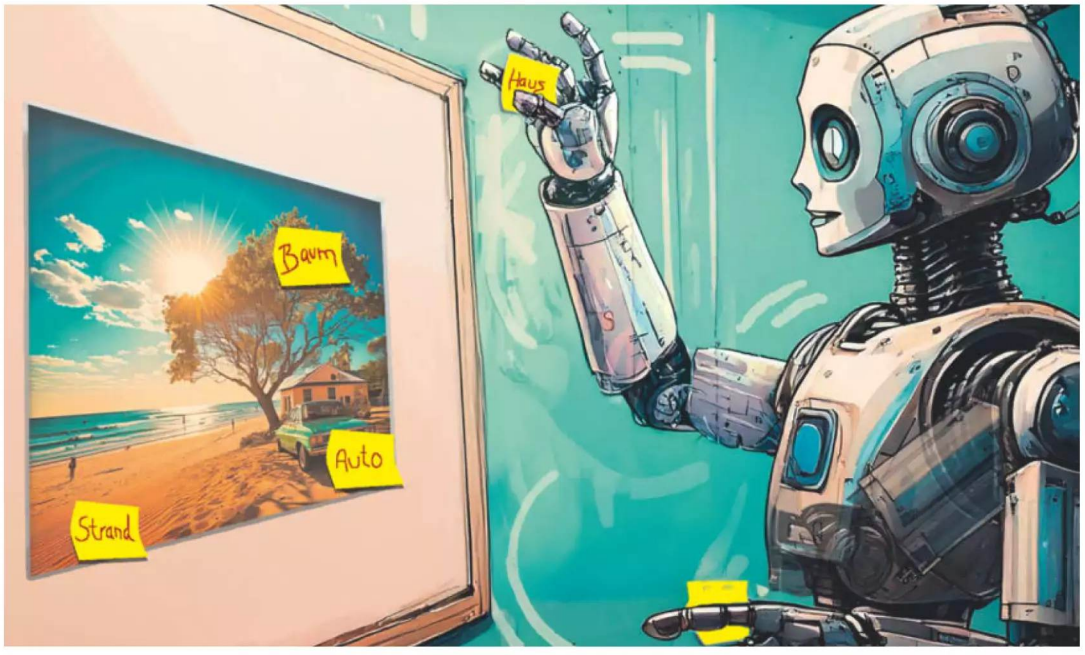
- .NET 10.0 im Überblick: SDK, Runtime und Basisklassen
- Neue Features in C# 14.0
- Webentwicklung in ASP.NET Core 10.0 und Blazor 10.0
- Neues beim OR-Mapping mit Entity Framework Core 10.0
- Veränderungen in WPF 10.0, Windows Forms 10.0 und WinUI 3
- Cross-Platform-Entwicklung mit .NET MAUI 10.0

Jetzt **Frühbucher-ticket** sichern!

Workshops am 20. / 25. / 27. November + 1. / 3. / 5. Dezember

net.bettercode.eu

Veranstalter iX Magazin für professionelle IT · dpunkt.verlag · in Kooperation mit www.IT-Visions.de Dr. Holger Schwichtenberg



(Bild: Jessica Nachtigall/KI/heise medien)

Bildverwalter mit KI-Verschlagwortung

Künstliche Intelligenz kann helfen, Bilder zu verwalten. Dafür muss man sie nicht den Clouddiensten von Adobe, Apple oder Google anvertrauen. Einige Bildverwaltungsprogramme nutzen lokale KI oder binden Sprachmodelle ein, um inhaltsbezogene Stichwörter zu vergeben.

Von **André Kramer**

Wer Fotos verschlagworten wollte, musste das lange Zeit manuell erledigen, und das war entweder aufwendig oder oberflächlich: „Italien“ ist eben weniger aussagekräftig als „Florenz; Uffizien; Statue; Skulptur; Renaissance; Michelangelo; David“. Entsprechend einfach oder schwierig findet man Motive anschließend über die Suchfunktion wieder. Umfangreiche Stichwortkataloge haben also ihren Sinn. Getreu dem Motto „Work smarter, not harder“ kann man solche Fleißarbeit dem Computer überlassen, beziehungsweise einer künstlichen Intelligenz.

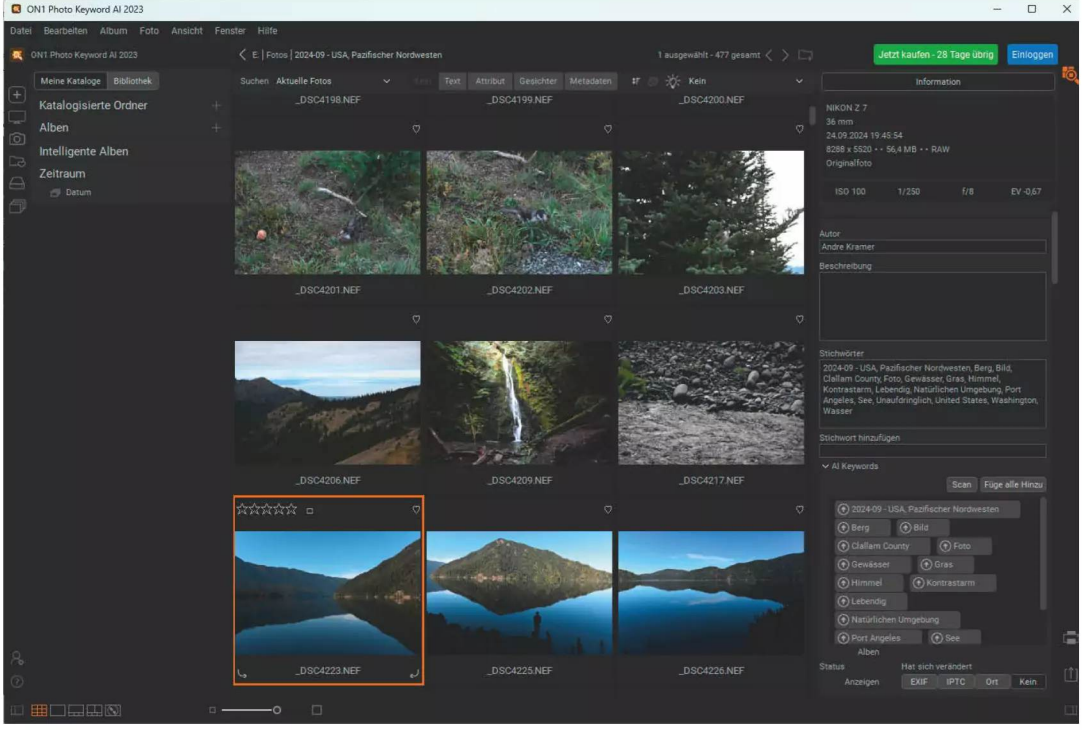
Cloudspeicherdienste machen es seit geraumer Zeit vor. Die KIs hinter den Foto-Apps von Apple und Google sowie Adobe Lightroom analysieren in die Cloud geladene Bilder im Hintergrund und vergeben zum Inhalt passende Stichwörter. In der Mobil-App oder einem Browser lässt sich der Medienbestand

nach Begriffen wie „Strand“, „Sonnenuntergang“, „Porträt“ oder „Schwarz-Weiß“ durchsuchen. Um die Nutzer bei der Stange zu halten, schlagen die Dienste außerdem regelmäßig Zusammenfassungen des Typs „Ein Tag am Strand“ oder „Festival-Stimmung“ vor. Auch die Open-Source-Software Immich, die unter anderem der in Europa gehostete Dienst PixelUnion nutzt, verschlagwortet mit KI. Das funktioniert mittlerweile bemerkenswert gut.

Clouddienste eignen sich allerdings nur für Smartphone-Fotos und solche, die der Nutzer zur Ansicht ins Netz exportiert hat. Es ist weder praktisch noch sinnvoll, 60 MByte große Raw-Fotos massenhaft noch vor dem Sichten und Bearbeiten ins Netz zu verfrachten, nur um sie anschließend zur Bearbeitung wieder auf den heimischen PC zu laden – von Datenschutzbedenken mal ganz abgesehen. Für die lokale Bildverwaltung braucht man also ein Programm mit eingebundener Stichwort-KI, das idealerweise auch die Themenfelder Kalendersuche, Gesichtserkennung sowie Geotagging abdeckt. Wir haben sechs Bildverwaltungsprogramme getestet, die automatisch Beschreibungen generieren: ACDSee Photo Studio Ultimate 2025, Excire Foto 2025 von PCR (Pattern Recognition Company), IMatch 2025.3 vom hessischen Entwickler Photools.com, Adobe Lightroom CC 8.4, Nitro 2025 von Gentleman Coders und ON1 Photo Raw 2025.

Verwalten mit künstlicher Intelligenz

Bei einigen Testkandidaten ist die Bildverwaltung nur ein Teilaspekt. Die Kernaufgabe von Lightroom CC ist das Entwickeln von Raw-Fotos. Das Adobe-Programm setzt für inhaltliche Verschlagwortung und Suche voraus, dass die Fotos in die Cloud verfrachtet werden. ACDSee Photo Studio und ON1 Photo Raw verschlagworten im Unterschied zu Lightroom lokal und verwalten Fotos daneben mit umfassenden Eingabemasken. Auch sie können in ihren Modulen zur Bildbearbeitung Raw-Fotos entwickeln. ON1 bietet alternativ das Tool Photo Keyword AI 2023 an, das lediglich KI-Stichwörter generiert. Der Testkandidat Excire Foto ging aus einem Forschungsprojekt der Uni Lübeck hervor und widmet sich



Der Fotoallrounder ON1 Photo Raw 2025 vergibt auch KI-Stichwörter. Die Funktion hat der Hersteller außerdem in das günstigere Programm ON1 Photo Keyword AI 2023 ausgelagert.

Sprachmodelle von OpenAI und Mistral in IMatch einbinden

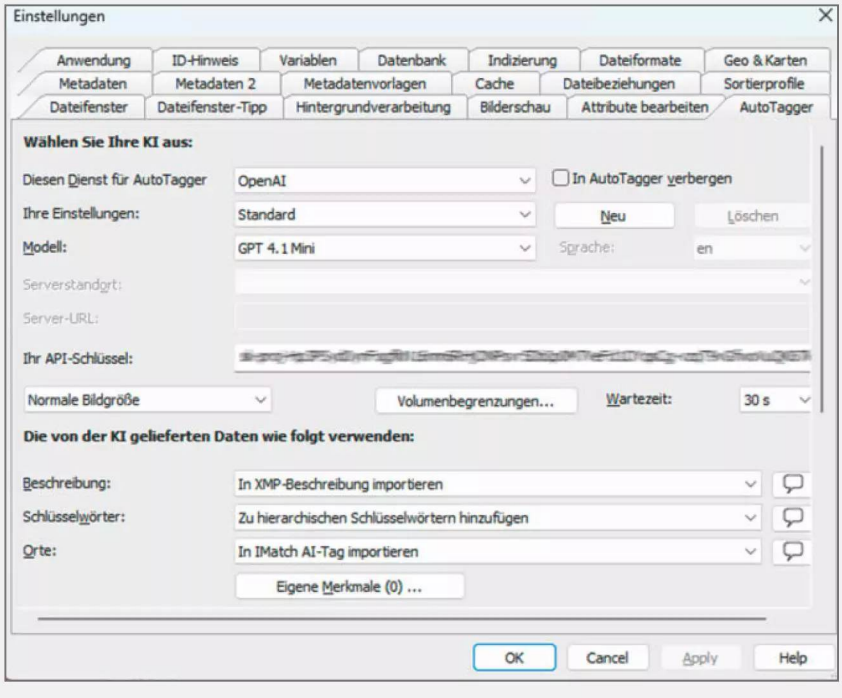
IMatch kommt mit Konfigurationen für die OpenAI-Modelle „gpt-4o-mini“ und „gpt-4o“ sowie für „pixtral-12b-latest“ von Mistral. Um sie nutzen zu können, muss man sich über die jeweilige Weboberfläche der externen Anbieter API-Keys besorgen und die Dienste bezahlen. Den Key trägt man in den Programmeinstellungen von IMatch unter „Bearbeiten/Einstellungen/AutoTagger“ ein.

Anschließend steht das Modell über „Befehle/Bild/AutoTagger“ zur Auswahl. Ein Klick auf „Run“ versieht ausgewählte Fotos mit einer Beschreibung. Über die Einstellungen-Schaltfläche lässt sich bestimmen, wo diese landen. So kann IMatch die Bildbeschreibung im Fließtext in die XMP-Beschreibung importieren. Es bietet sich an, für die per Semikolon getrennten Schlüsselwörter die Option „zu hierarchischen Schlüsselwörtern hinzufügen“ zu wählen.

Die Preise von GPT-4 in IMatch hängen vom Modell und vom Prompt ab. OpenAI berechnet bei GPT-4.1-mini aktuell 0,40 US-Dollar für eine Million Input-Token und 1,60 US-Dollar für eine Million Output-Token. Ein Token entspricht dabei etwa vier Zeichen. Bei jeweils 300 Input- sowie Output-Token pro Bild kann man demnach für 2 US-Dollar 3333 Bilder verschlagworten. Für 10 US-Dollar plus 1,90 US-Dollar Steuer verschlagwortet IMatch mit OpenAI-Key konservativ geschätzt etwa 15.000 Bilder.

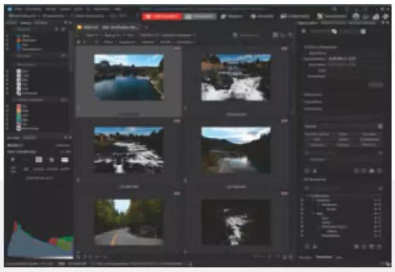
Prompts beeinflussen die Länge und den Stil von Beschreibungen. Mit IMatch-Variablen kann man in den Metadaten vorhandene Informationen wie Personen- oder Ortsangaben in Prompts verwenden. Statt eines Prompts wie „Describe this image“ können IMatch-Anwender zum Beispiel „Describe this image taken in {File.MD.city},{File.MD.country}“ verwenden. Der Kontext kann helfen, die Ergebnisse der KI entscheidend zu verbessern. Folgender Zusatz verbessert den Stil und produziert deutsche Stichwörter: „Describe this image in the style of a news headline. Use factual language. Respond in German language only.“

Mit sogenannten „IMatch Traits“ lassen sich Prompts und benutzerdefinierte Metadatenfelder kombinieren. Der Trait kann zum Beispiel „Contains Animals“ lauten und der Prompt dazu: „If there are animals visible in this image, respond with the text ‚Animals‘, else return an empty response“. Das Resultat kann AutoTagger im benutzerdefinierten Tag „AI.Animals“ speichern. Über die Variable {File.MD.AI.Animals} lässt es sich über die Filterfunktion nutzen.



Über Ollama und LM Studio können Nutzer verschiedene KI-Modelle auf ihren PCs laufen lassen und in IMatch einbinden, ohne dass dadurch Lizenzkosten oder Datenschutzprobleme entstehen. Beide Umgebungen unterstützen Llama, DeepSeek und Qwen. Ollama ermöglicht zudem Google Gemma 3 zu nutzen; LM Studio unterstützt auch Mistral.

In den Einstellungen von IMatch kann man für OpenAI GPT und Mistral API-Keys integrieren und festlegen, in welchen Datenfeldern die Beschreibungen landen sollen.



ACDSee Photo Studio 2025

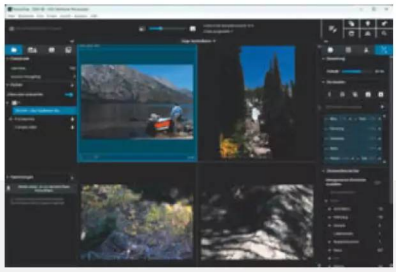
ACDSee gehört seit 1994 zu den Urgesteinen der Fotoprogramme. Zum kostenlosen Viewer und einigen Zwischenstufen gibt es die Ultimate-Version mit Bildverwaltung, Arbeitsbereichen zum Sichten und Vorführen, einem Raw-Entwickler und althergebrachter Bildbearbeitung. Darüber hinaus sind auch KI-Himmelersatz und -Objektauswahl an Bord. Der Clouddienst ACDSee 365 synchronisiert etwas schwerfällig zwischen PC und Smartphone.

Der Bildverwalter zeigt EXIF- und IPTC-Daten übersichtlich an und schreibt Änderungen auf Wunsch als XMP-Begleiter. Geotagging erledigt man durch Ziehen von Fotos auf eine integrierte Google-Karte.

Der Menüpunkt KI öffnet eine Gesichtserkennung, die Personen erkennt und zum Benennen vorschlägt. Sie funktioniert mittlerweile bemerkenswert gut. Außerdem enthält das Menü eine Stichwortanalyse, die deutlich hinzugelernt hat. Wenn man einen Ordner öffnet, kann man sie über besagtes Menü für alle gezeigten Fotos anstoßen. Die meisten anderen Programme erledigen das Bild für Bild.

Als einziges Programm im Test vergibt ACDSee hierarchische KI-Stichwörter. So enthält das Tag „Person“ das Stichwort Lächeln. Unter „Kleidung“ trägt ACDSee zum Beispiel Jeans und Schuhe ein. Die Stichwörter sind allesamt relevant und inhaltsbezogen. Stimmungen, Farben oder Kontraste erfasst ACDSee nicht. Über Checkboxen wählt man die Stichwörter aus und weist sie Fotos zu. Man kann auch pauschal alle übertragen. Ein Kontextbefehl bettet sie in das betreffende IPTC-Feld ein.

umfassende Bildverwaltung
praxisnahe KI-Schlagwörter
Preis: 179,99 €



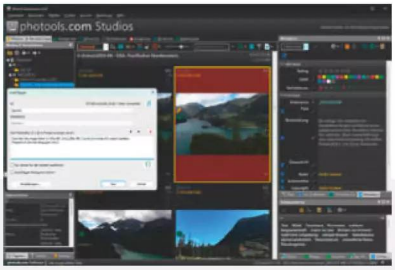
Excire Foto 2025

Excire Foto 2025 analysiert den Fotoinhalt und vergibt passende Schlagwörter. Das ältere Excire Search 2024 arbeitete zunächst nur als Plug-in für Lightroom Classic. Die Excire-Technik erkennt außerdem Gesichter, die es nach Alter, Geschlecht, Lächeln oder offenen Augen sortiert. Es sucht nach Bewertungen und Farbetiketten, findet Duplikate und ähnliche Fotos. Wenn die Fotos ein Geotag enthalten, kann man auf einer Karte zur Position einen Radius bestimmen, um Bilder aus der Umgebung aufzufinden. Metadaten lassen sich sowohl anzeigen als auch bearbeiten.

Nach Import eines Ordners beginnt Excire Foto mit der Analyse und versieht jeden Eintrag mit einer ganzen Reihe von Schlagwörtern. Das Programm arbeitet vollständig lokal, setzt also nach Registrierung der Software keine Onlineverbindung voraus. Die KI identifiziert hauptsächlich Motive wie „Elefant“, „Kirche“ oder „Wasserfall“. Hinzu kommen übergeordnete Genrebegriffe des Typs „Architektur“, „Reisen“ oder „Porträt“. Auch technische Details wie „ungesättigt“ oder „Komplementärfarben“ werden erfasst. Die Vorschläge markiert Excire mit einem Häkchen. Entfernt man es, landen diese Schlagwörter nicht in der Datenbank beziehungsweise den Metadaten.

Excire Foto macht seinen Job auf inhaltlicher Ebene bemerkenswert gut, und leistet sich nur leichte Schwächen beim Erfassen von Stimmungen. Es taggt Raw-Fotos beispielsweise oft als „kontrastarm“ und „ungesättigt“. Das ist nicht sachlich falsch, liegt bei Rohdaten aber in der Natur der Sache und sollte daher nicht in die Verschlagwortung eingehen.

konsequenter KI-Ansatz
kaum klassische Bildverwaltung
Preis: 199 €



IMatch 2025

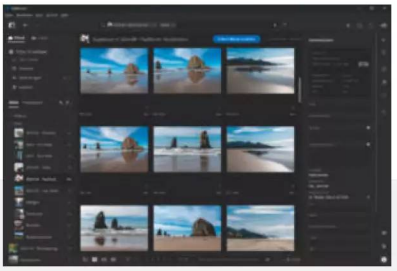
Die Bilddatenbank IMatch verwaltet Fotos mit einer Vielzahl eigener Markierungsarten. In deren Umgang liegt eine der Stärken des umfangreich konfigurierbaren Programms. Es schreibt saubere XMP-Begleiter und bearbeitet Content Credentials der Adobe Content Authenticity Initiative (CAI).

Praktische Tastenkürzel öffnen die Vergleichs- oder Vollbildansicht. Das Filtertütensymbol in der Titelzeile ruft einen Schnellfilter auf. In einer Zeile unterhalb der Symbolleiste kann man nach Bewertungen und Etiketten filtern. Der gelungene Vergleichsmodus zeigt mehrere Fotos auch in Zoomansicht. Verändert man mit der Maus den Ausschnitt eines Bilds, überträgt IMatch diesen auf alle geöffneten Bilder.

Die Gesichtserkennung gruppiert Personen, sodass man diese nur noch benennen muss. Einträge für Familie, Freunde, Kollegen oder Kunden sortieren sie. Das integrierte Geotagging-Modul bindet Onlinekarten von Google, Bing, Here und OpenStreetMap ein. Mithilfe der Aufnahmedaten lassen sich dynamische Alben als sogenannte „Events“ anlegen. Dafür gibt man Start- und Enddatum ein.

Neu in IMatch 2025 ist der AutoTagger 2.0. Er bindet kostenpflichtige API-Keys von OpenAI und Mistral ein, um passend zu ausgewählten Fotos Bildbeschreibungen und Schlüsselwörter zu generieren (siehe Kasten). Mit Prompts kann man angeben, wo das Foto entstand und wie sowie in welcher Sprache die KI den Inhalt beschreiben soll. Das ist mit etwas Aufwand verbunden, aber mit keinem Programm gelingen genauere und bessere Stichwörter.

großer Funktionsumfang
hoher Konfigurationsaufwand
Preis: 135 €



Lightroom CC 8.4

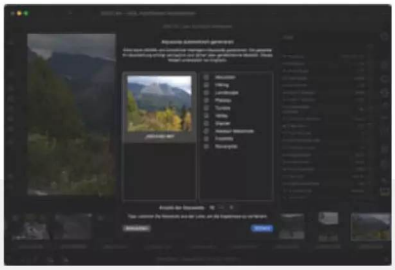
Adobe setzt bei seinem Raw-Entwickler Lightroom CC ganz auf die Cloud. Das bereits 2017 eingeführte Programm soll Lightroom Classic ersetzen, das mit einer lokalen Fotobibliothek arbeitet. Hinsichtlich der Fotoentwicklung gleichen sich die beiden Programme. Die KI-Auswahl von Motiven und dem Himmel arbeitet auch bei Lightroom Classic schon in der Cloud. Zwar kann man den Classic-Katalog in Lightroom CC importieren, braucht dann aber Cloudspeicher, den Adobe sich bezahlen lässt. Im Abo für 14,49 Euro pro Monat ist ein Terabyte Cloudspeicher enthalten.

Große Unterschiede gibt es bei der Bildverwaltung. Nur Lightroom Classic enthält ein Kartenmodul fürs Geotagging und einen umfangreichen IPTC-Editor. Nur in Lightroom Classic kann man ortsbezogen suchen. Dessen Gesichtserkennung arbeitet wiederum schwerfälliger als das bessere System von Lightroom CC.

Lightroom CC analysiert im Hintergrund alle Bilder in der Cloud und generiert KI-Beschreibungen, die sich aber nicht im Klartext anzeigen oder als Metadaten ins Foto exportieren lassen. In exportierten XMP-Begleitern tauchen keine Stichwörter auf. Adobe bietet hier ein hermetisch abgeschlossenes System. Software von Drittanbietern lässt sich nicht in den Workflow integrieren.

Auf der anderen Seite kann man sich im Adobe-System auch wohlfühlen: Das Desktop-Lightroom synchronisiert sich nahtlos mit den Programmvarianten für Smartphone und Tablet. Über die Suchzeile lässt sich im Katalog intuitiv recherchieren, und generell findet man das Gesuchte auch.

gut bedienbare KI-Verwaltung
geschlossenes Adobe-System
Preis: 14,49 € pro Monat



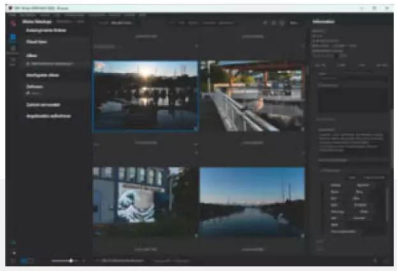
Nitro 2025

Nitro ist für macOS erhältlich und deckt den gesamten Raw-Workflow ab. Nach Import kann man mit bis zu fünf Sternen bewerten sowie als angenommen oder abgelehnt markieren. Zusätzlich zu Übersichts-, Einzel- und Vollbildansicht gibt es auch einen exzellenten Vergleichsmodus, der Zoom und Bildausschnitt synchron für alle gezeigten Bilder vollzieht.

In der Einzelbildansicht kann man über eine Seitenleiste verschiedene Paletten aufrufen. Der Entwicklungsmodus enthält Raw-Werkzeuge für Grundeinstellungen, HSL-Bearbeitung (Farbton, Sättigung und Luminanz), Objektiv- und Perspektivkorrektur, selektive Arbeit mit Masken sowie Retusche. Weitere Paletten blenden verschiedene Metadaten ein. Nitro bringt einige hierarchische Standard-Schlagwörter mit, die vom Hersteller übersetzt wurden. Ältere Versionen enthielten Fehler. Bestandskunden müssen die alte Keyword-Liste manuell entfernen, um die neue zu erhalten.

Nitro nutzt eine Apple-KI mit 100 Millionen Parametern, um Fotos lokal zu verschlagworten. Mit dem Befehl „Metadaten/Automatische Keywords“ ruft das Programm einen Dialog auf, der pro Bild je nach Auswahl drei bis zehn Schlagwörter erzeugt; leider ausschließlich englischsprachige. Meistens passt nur ein Teil der Vorschläge. Oft rät die KI den Aufnahmeort und ergänzt beispielsweise neben „Lake“ und „Reflection“ auch den Begriff „Newfoundland“, obwohl ein Geotag ausweist, dass das Bild nicht in Kanada entstand. Eine Straße taggte es als „Highway“ (Schnellstraße) und „Bridle Road“ (Reit- und Wanderweg). Beides zusammen geht nicht.

guter Vergleichsmodus
nur englische Stichwörter
Preis: 99,99 € oder 34,99 €pro Jahr



ON1 Photo Raw 2025

ON1 hat früher Photoshop-Plug-ins hergestellt und später alles zum Raw-Entwickler ON1 Photo Raw verheiratet. Dieser bietet unter anderem Foto-Effekte, Filmsimulation und Porträtbearbeitung. Seit Version 2024 ist auch eine Stichwort-KI an Bord. ON1 Photo Raw 2025 kostet rund 100 Euro; zusammen mit Plug-ins und Clouddiensten etwas mehr. Den KI-Stichwortdienst hat der Hersteller ausgekoppelt: ON1 Photo Keyword AI 2023.5 kostet 74,85 Euro. Bis Redaktionsschluss gab es auf alle genannten Preise 60 Prozent Nachlass.

Den Auftrag, KI-Schlagwörter zu generieren, hat ON1 unaufdringlich und leicht auffindbar im Metadatenbereich untergebracht. Unter den Textfeldern für Autor und Beschreibung findet sich eines für die Stichwörter. Die kann man entweder manuell hinzufügen oder über einen Klick auf „Scan“ von der KI ergänzen lassen. Voreingestellt ist ein Stichwort mit dem Ordnernamen. Das Programm fügt nach Klick etwa ein Dutzend Wörter hinzu.

Die Qualität der Schlagwörter ist mäßig. Nahezu jedes Foto enthält das unsägliche „kontrastarm“. Die meisten Fotos taggt ON1 zudem als „Bild“ – ach was. Sobald Menschen zu sehen sind, geht es los: Erwachsene, männliche Person, menschlich, menschliche Aktion, menschlicher Kopf, Person, Personen, Oberbekleidung – deutlich zu viel. „Fahrzeug“ ist fast immer dabei, auch wenn nur ein Kanu zu sehen ist. Immer wieder findet sich neben „unaufdringlich“ auch „stummgeschaltet“, eine wohl verunglückte Übersetzung für „muted“ (gedämpft).

gut integrierte KI-Verwaltung
Mängel bei Stichwörtern
Preis: 106,93 € (Komplettpaket 213,89 €)

Bildverwaltung mit künstlicher Intelligenz

Produkt	ACDSee Photo Studio Ultimate 2025	Excire Foto 2025 (4.0)
Hersteller, URL	ACD Systems, acdsee.com	Pattern Recognition Company, excire.com
Systemanforderungen	Windows ab 10 (64 Bit)	Windows ab 10 (64 Bit), macOS ab 11
Sprache	Deutsch	Deutsch
Import		
Arbeitsweise	Ordner öffnen / katalogisieren	Import in Datenbank
Formate: JPEG, PNG, TIFF / Raw	✓ / ✓	✓ / ✓
JPG2000 / HEIF / WebP	– / ✓ / ✓	– / ✓ / –
MOV / MP4	✓ / ✓	✓ / ✓
Import im Hintergrund / aus Unterordnern	✓ / ✓	✓ / ✓
ICC-Farbverwaltung	✓	–
Metadaten: dateiintern / XMP-Begleiter	✓ / ✓	✓ / ✓
Darstellung		
Bildschirmfüllend / pixelgenau (100 %)	✓ / ✓	✓ / ✓
Details in Vollbild / mit Dateneingabe	✓ / ✓	✓ / ✓
Schnellkollektion	✓ (Auswahlkorb)	✓ (Flagge)
Ordner / Kalender / Alben	✓ / ✓ / ✓ (Katalog)	✓ / ✓ / ✓ (Sammlung)
Anzeigen: zuletzt verwendet	–	✓ (letzter Import)
Verwaltung		
Gesichtserkennung	✓	✓
Geotagging	✓ (Fenster/Karte)	✓ (Koordinateneingabe)
KI-Schlagworte	✓ (hierarchisch)	✓
Bewertung / Farbetikett	✓ / ✓	✓ / ✓
Kategorien / hierarchisch / einbetten	✓ / ✓ / ✓	✓ / ✓ / ✓
XMP-Begleiter exportieren	✓ (automatisch)	✓ (manuell)
EXIF-Datum setzen / verschieben	✓ / ✓ (numerisch)	– / –
IPTC-Eingabemaske / konfigurierbar	✓ / ✓	✓ / ✓
Tastenkürzel für 5 Sterne	Strg+5	5
Suche		
komplexe Suche / Schnellsuche	✓ / ✓	– / ✓
Filter: Bewertung / Farbetikett	✓ / ✓	✓ / ✓
Filter: Gesichter / Ort	✓ / ✓ (Schwenk, Zoom)	✓ / ✓ (Ort, Radius)
Duplikate / Ähnlichkeit	✓ / –	✓ / ✓
Ausgabe und Bearbeitung		
Bildbearbeitung	umfassender Raw-Entwickler	–
Bildexport: umbenennen / skalieren / Metadaten entfernen / Wasserzeichen	✓ / ✓ / ✓ / ✓ (Aktionen)	✓ / ✓ / ✓ / –
Metadatenvorlagen speichern	✓	✓
Cloud-Anbindung	✓ (ACDSee 365, OneDrive)	✓ (Google Drive, Dropbox)
Begleit-App	ACDSee Mobile Sync	–
Diashow	Übergangseffekte, Schwenk und Zoom, Musik, Texttitel, nur Ansicht	Anzeigedauer, Übergangsdauer, -effekte, Hintergrundfarbe, Schleife
Bewertung		
Bedienung	○	⊕
KI-Bildverwaltung	⊕⊕	⊕⊕
Gesichtserkennung	⊕	⊕⊕
Anzeige und sichten	⊕	○
Bearbeiten und entwickeln	○	nicht verfügbar
Preis	179,99 €	199 €

⊕⊕ sehr gut ⊕ gut ○ zufriedenstellend ⊖ schlecht ⊖⊖ sehr schlecht ✓ vorhanden – nicht vorhanden k. A. keine Angabe

	IMatch 2025.3	Lightroom CC 8.4	Nitro 2025	ON1 Photo Raw 2025
	Photools.com, photools.com	Adobe, adobe.com	Gentleman Coders, gentlemencoders.com	ON1, on1.com
	Windows ab 10 (64 Bit)	Windows ab 10 (64 Bit), macOS ab 13.1	macOS ab 13.3, iOS/iPadOS ab 16.4	Window ab 10 (64 Bit), macOS ab 10.15
	Deutsch	Deutsch	Deutsch	Deutsch
	Import in Datenbank	Import in Datenbank	Import in Datenbank	Ordner öffnen / katalogisieren
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ / ✓ / ✓	– / ✓ / –	✓ / ✓ / ✓	– / ✓ / –
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ (Miniaturen optional)	✓	✓	✓
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ / ✓	✓ / ✓	✓ / ✓	– / –
	✓ (z. B. Flagge)	✓ (Flagge)	✓ (Flagge)	✓ („Liked“)
	✓ / ✓ / ✓ (Kollektionen)	✓ / – / ✓ (Sammlung)	✓ / – / ✓ (Ordner)	✓ / ✓ / ✓
	✓ (kürzlich aktualisiert)	✓ (letzter Import)	–	✓ (letzter Import)
	✓ (Familie, Gruppen)	✓	–	✓
	✓ (auf Karte platzieren)	– (nur LR Classic)	– (angekündigt)	✓ (auf Karte platzieren)
	✓ (mit externen LLMs)	✓ (versteckt)	✓ (bis zu zehn pro Foto)	✓
	✓ / ✓	✓ / ✓	✓ / – (nur Flagge)	✓ / ✓
	✓ / ✓ / ✓	✓ / – / ✓	✓ / ✓ / ✓	✓ / ✓ / ✓
	✓ (auto. oder manuell)	✓ (über Export/Original)	✓ (automatisch)	✓ (auto. oder manuell)
	✓ / ✓ (numerisch)	✓ / ✓ (stundenweise)	– / –	✓ / ✓ (stundenweise)
	✓ / ✓	✓ / –	– / –	✓ / ✓
	5	5	5	5
	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
	✓ / ✓	✓ / ✓	✓ / – (nur Flagge)	✓ / ✓
	✓ / ✓ (Ort, Radius)	✓ / ✓ (Schwenk, Zoom)	– / –	✓ / ✓ (über IPTC-Ort)
	✓ / ✓ (nach Form, Farbe)	– / –	– / –	– / –
	–	umfassender Raw-Entwickler	umfassender Raw-Entwickler	Raw-Entwickler, Effekte, Porträt, KI-Filter
	✓ / ✓ / ✓ / ✓ (Stapelverarbeitung)	✓ / ✓ / ✓ / ✓	✓ (manuell) / ✓ / ✓ / ✓ (umfangreich)	✓ / ✓ / ✓ / ✓
	✓	✓	–	✓
	–	✓ (Adobe Creative Cloud)	–	✓ (ON1 Cloud Sync)
	–	Lightroom CC	Nitro für iOS, iPadOS	ON1 Photo RAW for Mobile
	Anzeigedauer, Übergangseffekte, Hintergrundfarbe, Schleife, Text einblenden	Wechsel über Pfeiltasten (Anzeigedauer einstellbar)	Wechsel über Pfeiltasten	Wechsel über Pfeiltasten
	⊖⊖	⊕⊕	⊕	○
	⊕	⊕	⊖	○
	⊕	⊕	nicht verfügbar	⊖
	○	⊕	⊕⊕	⊕
	nicht verfügbar	⊕⊕	○	⊕
	135 €	14,49 € pro Monat	99,99 € oder 34,99 € pro Jahr	106,93 € (Komplettpaket 213,89 €)

gänzlich der KI-Bildverwaltung, ebenfalls lokal. Alle bisher genannten Programme sind für Windows und macOS erhältlich.

Die Bilddatenbank IMatch steht nur für Windows zur Verfügung und kann mit etwas Konfigurationsaufwand verschiedene KI-Modelle über API-Schlüssel einbinden. Damit lassen sich genauere und flexiblere Beschreibungen generieren als mit fertig konfigurierten Werkzeugen, man muss sich aber auch etwas einarbeiten (siehe Kasten). Das macOS-Tool Nitro verschlagwortet mit der im System verankerten Apple-KI. Der Nitro-Autor Nik Bhatt, alias Gentlemen Coders, hat früher als Entwickler bei Apple an der mittlerweile nicht mehr vertriebenen Fotosoftware Aperture gearbeitet.

Künstliche Intelligenz hilft in einem anderen Feld schon länger bei der Bildverwaltung. Paradebeispiel und prädestiniert für Mustererkennung mit maschinellem Lernen ist die Gesichtserkennung. Besonders gut machen das Dienste wie Google Fotos vor, aber auch ACDSee, Excire Foto, IMatch und Lightroom CC gruppieren Fotos zuverlässig nach Personen, sodass man denen nur noch Namen geben muss. Lediglich Sonnenbrillen und großer Altersunterschied beispielsweise bei aktuellen und Kinderfotos derselben Personen bereiten ihnen noch Schwierigkeiten.

Austausch mit anderen Programmen

Die Rolle als Raw-Komplettlösung muss man nicht akzeptieren. Capture One und DxO PhotoLab sind zum Beispiel exzellente Raw-Entwickler, bei der Bildverwaltung aber eher schwach aufgestellt. Auf der anderen Seite ist ACDSee eine starke Bildverwaltung, aber nicht der beste Raw-Entwickler. Darüber hinaus gibt es von der Fotoentwicklung unabhängige Einsatzgebiete, um große Bildersammlungen zu betreuen, beispielsweise in Unternehmen.

Für einen reibungslosen Workflow zwischen Programmen verschiedener Hersteller spielen Metadatenstandards und deren saubere Umsetzung eine große Rolle. Kameras halten in den EXIF-Daten Informationen zu Kameramodell, Objektiv, Brennweite, Empfindlichkeit, Belichtungszeit und Aufnahmedatum fest. Letzteres lässt sich in einer Bildverwaltung ändern, der Rest ist festgeschrieben. Über den IPTC-IIM-Standard (kurz IPTC) vom International Press Telecommunications Council lassen sich inhaltliche Informationen wie Name und Adresse des Fotografen sowie Titel, Beschreibung und Stichwörter zum Inhalt festhalten. Er ist auf die Bedürfnisse der Presse ausgerichtet, hat sich aber allgemein durchgesetzt.

Diese und weitere Daten, beispielsweise Entwicklungseinstellungen von Lightroom, lassen sich als XMP-Begleiter (Extensible Metadata Platform) für jedes Foto exportieren. Die Begleitdateien tragen dem Umstand Rechnung, dass Raw-Dateien in der Regel unverändert bleiben. Der von Adobe im Jahr 2001 veröffentlichte Standard nutzt die formale Sprache RDF (Resource Description Framework). XMP-Dateien speichern die Metadaten in einem XML-Dialekt und sind im Klartext lesbar. Alle Testkandidaten bis auf Lightroom hinterlegen ihre KI-generierten Stichwörter entweder automatisch oder auf Wunsch in XMP-Begleitern. So lassen sie sich in eine andere Software importieren.

Fazit

Kunden von Lightroom CC bekommen das Komplettpaket aus KI-gestützter Bildverwaltung und Raw-Entwickler für Windows, macOS, Android und iOS. Das funktioniert wunderbar und nahtlos zwischen Desktop- und Mobil-Apps. Adobe verkauft aber ein geschlossenes System als Blackbox. Die KI-Beschreibungen bekommt man nicht aus dem Programm heraus. Man kann nicht nachvollziehen oder ändern, was die KI macht, und keine Programme von Drittanbietern einbinden.

Die Bildverwaltung IMatch ist dagegen maximal konfigurierbar. Wer bereit ist, sich einzuarbeiten, kann die Oberfläche so konfigurieren, wie es dem persönlichen Nutzungsszenario entspricht. Über API-Keys kann man Sprachmodelle von OpenAI und Mistral einbinden und mit Prompts detailliert steuern. Mit keinem Programm im Testfeld entstanden bessere Beschreibungen. Daneben entstehen überschaubare Zusatzkosten für die KI-Dienste wie GPT-4.

Die übrigen vier Kandidaten bewegen sich zwischen diesen Polen. Alle verschlagworten lokal. Excire Foto widmet sich ausschließlich der KI-Verwaltung inklusive guter Gesichtserkennung. ACDSee, Nitro und ON1 Photo Raw haben außerdem einen Raw-Entwickler an Bord. Nitro arbeitet leider nur mit englischsprachigen Stichwörtern und erkennt keine Personen. Das ON1-Programm vergibt etliche redundante und nichtssagende Stichwörter wie „Bild“. Auch Excire Foto findet neben vielen relevanten immer auch einige nichtssagende Begriffe, macht seine Arbeit aber im Großen und Ganzen gut. ACDSee erzeugt einen sauberen, hierarchischen Katalog als Zusatzfunktion einer auch sonst soliden Bildverwaltung. (akr) c't

RAUS AUS DEN US-CLOUDS!

Wie Sie sich selbst aus der Abhängigkeit großer Tech-Konzerne befreien und wieder souverän über Ihre Daten und Dienste verfügen, zeigt Ihnen die c't-Redaktion anhand zahlreicher Beispiele in diesem Sonderheft:

- Gründe für den Cloud-Ausstieg
- Alternativen zu US-Clouddiensten
- Admin-Wissen für die private Cloud
- Nextcloud: Wieso, weshalb, warum
- Lohnenswerte Self-Hosting-Projekte

JETZT UMSTEIGEN!

shop.heise.de/ct-digital-souveraen

Generell portofreie Lieferung für Heise Medien- oder Maker Media Zeitschriften-Abonnenten oder ab einem Einkaufswert von 20 € (innerhalb Deutschlands). Nur solange der Vorrat reicht. Preisänderungen vorbehalten.

heise shop

Bild: Martina Bruns/KI/heise medien

Wie Big-Tech die Wirtschaft bedroht

US-Konzerne wie Google und Amazon kontrollieren weltweit Märkte und Datenflüsse. Die „glorreichen Sieben“ setzen Marktmechanismen außer Kraft, beeinflussen demokratische Prozesse und pfeifen aufs Klima, wenn sie ihre gigantischen Serverzentren hochziehen. Mit künstlicher Intelligenz verschaffen sie sich nun ein neues Machtinstrument, das ihre Dominanz weiter absichert.

Von **Hartmut Gieselmann**

Technik hat etwas Faszinierendes: Viele bekommen leuchtende Augen, wenn ein neuer Prozessor den PC beschleunigt, Computerspiele fotorealistische Welten zaubern oder eine künstliche Intelligenz die Antwort auf jede Frage der Welt zu kennen scheint. Für deren Entwicklung schien die Marktwirtschaft lange das ideale Umfeld zu sein: Ideen und Produkte treten in freien Wettbewerb, die besten setzen sich durch und erleichtern das Leben aller – so jedenfalls die Idealvorstellung

von Adam Smith, der am Vorabend der Französischen Revolution die Grundlagen der kapitalistischen Ordnung entwarf.

Mehr als zweihundert Jahre später tritt der Kapitalismus jedoch in eine neue Phase. Statt Fabrikhallen und Eisenbahnnetzen entstehen heute gigantische Rechenzentren und immer schnellere Datenautobahnen. Doch diese globale Infrastruktur wird nicht von einer Vielzahl konkurrierender Unternehmen errichtet, sondern von einer Handvoll US-Giganten dominiert. Zu diesen „Magnificent Seven" (M7), wie sie an den Börsen heißen, gehören Nvidia, Microsoft, Apple, Amazon, Alphabet, Meta sowie das Firmenkonglomerat von Elon Musk um Tesla und SpaceX. Gemeinsam erreichen sie inzwischen einen Marktwert von 18 Billionen Euro.

Damit scheint ein ökonomischer Kipppunkt erreicht: Eine kleine Gruppe von Konzernen verfügt – zumindest nach aktuellem Handelswert ihrer Anteilsscheine – über eine Kapitalmacht, die der Jahreswirtschaftsleistung aller Unternehmen und knapp 450 Millionen Bürger der Europäischen Union entspricht. Auf den folgenden Seiten analysieren wir, wie es zu diesem kometenhaften Aufstieg kommen konnte und welche Rolle künstliche Intelligenz dabei spielt. Außerdem beleuchten wir die ökologischen Auswirkungen der gigantischen Serverzentren auf die Umwelt (ab Seite 168) und welche Möglichkeiten Initiativen in Europa haben, sich von dieser Machtkonzentration zu emanzipieren (ab Seite 176).

Der Kapitalismus geht in Rente

Schlüssige Analysen, wie die M7 derartig aufsteigen konnten, liefern die beiden Ökonomieprofessoren Cédric Durand aus Frankreich und Yanis Varoufakis aus Griechenland. In seinem Buch „Technofeudalismus – Was den Kapitalismus tötete" macht Varoufakis die Austeritätspolitik der Zentralbanken als einen wichtigen Faktor aus [1]. Demzufolge pumpten Zentralbanken seit der Finanzkrise im Jahr 2008 Billionen billigen Geldes in die Märkte. Bei den M7 wirkte das wie Dünger: Sie kauften eigene Aktien zurück, trieben die Kurse künstlich nach oben und sammelten Kapital an, das sie in immer größere Plattformen steckten.

Fast jedes Mitglied der M7 hat in seinem Bereich als sogenannter Gatekeeper eine marktbeherrschende Stellung erreicht. Mit Marktanteilen von 40 Prozent und mehr können sie Preise diktieren und Konkurrenten aus dem Feld schlagen. Damit gleichen sie laut Varoufakis den Feudalherren vergangener Zeiten: Diese ließen Bauern Abgaben zahlen, sobald sie das Land bearbeiteten – sie schöpften bestehende Werte ab, ohne durch Arbeit selbst neue zu schaffen.

Die M7 kassieren heute nach demselben Prinzip: nicht auf Äckern, sondern auf digitalen Plattformen, wo jeder Klick, jedes Abo und jede Rechenstunde eine Rente für die Konzerne abwirft. Ökonomen meinen damit keine Altersversorgung, sondern Gewinne, die auf der Kontrolle von Zugängen beruhen, von denen Nutzer, Entwickler und Unternehmen immer stärker abhängig werden [2]. Für die Konzerne ist das lukrativer, als neue Werte zu produzieren – Adam Smiths Traum vom „Wohlstand der Nationen" geht damit zu Ende [3]. Wie solche Mechanismen langfristig zu extremer Vermögenskonzentration führen, hat Thomas Piketty in „Das Kapital im 21. Jahrhundert" gezeigt [4].

Feudaler Marktplatz

Wie der Technofeudalismus funktioniert, zeigt das Beispiel Amazon: Der US-Konzern kontrolliert inzwischen allein in Deutschland rund 60 Prozent des Onlinehandels. Auf dem Marketplace wählen nicht die Nutzer frei das beste Produkt, sondern der Algorithmus von Amazon: Unabhängige Studien etwa der Harvard-Universität (siehe ct.de/wuhg) zeigen, dass Amazon-eigene sowie Prime-Produkte zuverlässig höher platziert werden als gleichwertige Artikel Dritter. Diese algorithmische Bevorzugung ist keine Ausnahme, sondern Teil der ökonomischen Struktur: Sichtbarkeit wird zur Ressource, die Amazon über seine integrierten Marken monopolisiert – auf Kosten von Wettbewerb und Qualitätskriterien.

So fließt bei jedem Kauf ein großer Teil des Preises nicht an den Hersteller, sondern an Amazon: durchschnittlich 15 Prozent Vermittlungsgebühr, in manchen Fällen bis zu 45 Prozent. Sichtbarkeit im Prime-Programm kostet weitere 20 bis 35 Prozent Fulfillment-Gebühr. Selbst wer den Versand eigenständig organisiert, muss zwei Prozent „Prime-Gebühr" abtreten. Werbung beim „Prime Day" schlägt zusätzlich mit bis zu 500 Euro pro Platz im digitalen Schaufenster zu Buche. In der Summe wandern so oft 30 bis 50 Prozent des Kaufpreises direkt an Amazon – Kosten, die vor den Kunden verborgen bleiben, von den Händlern aber eingepreist werden müssen.

Damit sitzt Amazon an der entscheidenden Schnittstelle zwischen Anbietern und Käufern. Der Konzern kontrolliert Zahlungsverkehr und Lieferung

und verhindert, dass Händler ein direktes Verhältnis zu ihren Kunden aufbauen. Es ist, als ob ein Wochenmarkt nicht mehr auf dem Dorfplatz stünde, sondern auf einem Privatgrundstück mit Einlasskontrolle.

Dank des World Wide Web ist Amazon aber nicht bloß der Grundherr eines Dorfmarktes, sondern in allen westlichen Industrienationen der größte digitale Feudalherr. Seine Plattform hat ganze Kaufhausketten aus dem Markt gedrängt und Fußgängerzonen entvölkert. 2024 überstiegen die weltweiten Umsätze des Konzerns mit 593 Milliarden Euro sogar den Bundeshaushalt Deutschlands von 477 Milliarden Euro.

Digitale Rentenplattformen

Eng damit verknüpft ist der Begriff des Plattformkapitalismus, den der britische Politökonom Nick Srnicek 2017 untersuchte [5]. Apple und Google kontrollieren mit ihren App-Stores den weltweiten Markt für mobile Anwendungen. In Europa liegt der Anteil von Android-Smartphones derzeit bei rund zwei Dritteln, iPhones bei etwa einem Drittel.

Für Srnicek sind Apple und Google weder klassische Händler noch Produzenten, sondern Betreiber einer Infrastruktur, auf der andere arbeiten müssen. Je mehr Entwickler und Nutzer sich dort versammeln, desto schwerer wird es für alle Beteiligten, die Plattform zu verlassen. Aus dieser wachsenden Abhängigkeit entsteht die eigentliche Macht – und die Möglichkeit, Gebühren nach Belieben zu erheben.

Seit der Eröffnung der Stores im Jahr 2008 erheben beide Konzerne auf jeden App-Kauf eine Gebühr. Über ein Jahrzehnt lang lag diese bei 30 Prozent. Erst nach massiven Beschwerden großer Anbieter wie Spotify, Epic Games oder Tinder und unter wachsendem politischen Druck senkten Apple und Google 2021 die Abgaben für kleinere Entwickler mit weniger als einer Million US-Dollar Jahresumsatz auf 15 Prozent. Am Grundmechanismus hat sich jedoch nichts geändert: Wer seine App einer breiten Öffentlichkeit zugänglich machen will, kommt an den Stores nicht vorbei.

Der deutsche Soziologe Philipp Staab beschreibt dieses Prinzip als Kern des „digitalen Kapitalismus“ [6]: Digitale Güter wie Apps, Software oder Musik lassen sich eigentlich beliebig oft und fast ohne Kosten vervielfältigen. Doch Plattformbetreiber wie Apple und Google schotten ihre Ökosysteme gezielt ab, um künstliche Knappheit zu erzeugen. Sichtbarkeit, Vertriebswege und Zugang zu den Nutzern werden so zur knappen Ressource – und damit zur Einnahmequelle für die Plattformen.

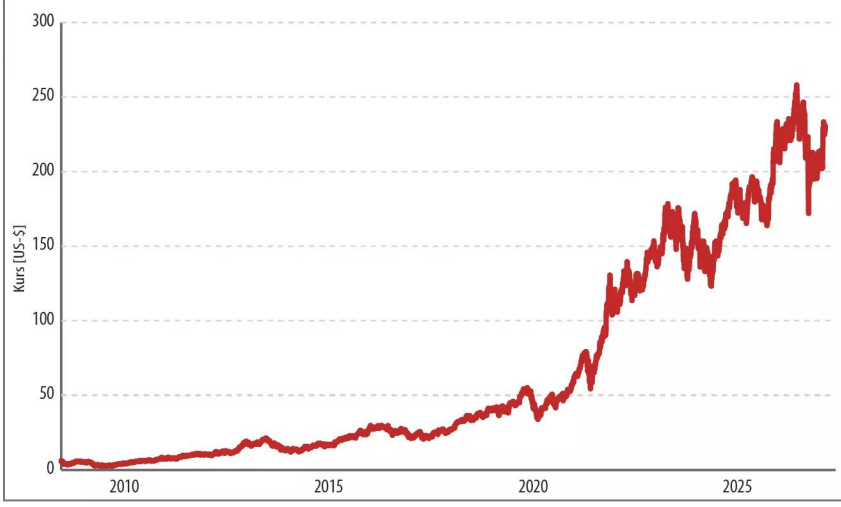
Die Regulierung durch den Digital Markets Act von 2022 schreibt zwar vor, dass Apple und Google auch alternative App-Stores zulassen müssen. Doch beide Unternehmen stellen immer wieder neue technische Hürden auf, um ihre Dominanz zu wahren. Apple beharrt darauf, jede App weiterhin selbst prüfen zu dürfen, und erhebt zusätzliche Gebühren für Anwendungen mit mehr als einer Million Downloads. Unter Android lassen sich Apps zwar auch per Sideload durch den Download einer APK-Datei installieren, doch ohne den Play Store verlieren sie den Zugang zu den Google-Diensten und deren Standort-, Log-in- und Bezahlfunktionen sowie den Push-Nachrichten, die für viele Anwendungen unverzichtbar sind.

Bild: Luca Bruno/AP/dpa

Bezos in Venedig: Die private Hochzeit des Amazon-Chefs kostete zwischen 40 und 48 Millionen Euro – nicht eingerechnet die Kosten für den italienischen Staat, der große Teile der Stadt von Polizei und Militär aufgrund von Protesten absperren ließ.

Apple Inc. (AAPL)

Von 2,50 auf über 200 US-Dollar: Apples Kurs-Rallye seit der Finanzkrise 2008 ist auch das Ergebnis einer gigantischen Rückkauforgie seit 2012 im Volumen von 650 Milliarden US-Dollar – ermöglicht durch das billige Geld der Fed.



Quelle: Macrotrends, www.macrotrends.net

Daten- und Aufmerksamkeitskontrolle

Außer dem Waren- und Onlinehandel privatisierten die M7 über ihre Social-Media-Plattformen die weltweite Kommunikation und Informationsverbreitung: Meta kontrolliert WhatsApp, Instagram und Facebook, Alphabet die Internetsuche sowie YouTube, Elon Musk hat sich Twitter einverleibt und in X umbenannt.

Welche Rolle die gesammelten Daten bei der Überwachung und Rentenabschöpfung spielen, beschreibt die US-Sozialwissenschaftlerin Shoshana Zuboff in ihrem Buch „Das Zeitalter des Überwachungskapitalismus“ [7]. Sie zeigt, wie die Konzerne persönliche Informationen in handelbare Vorhersagen über das Verhalten ihrer Nutzer verwandeln – ein Rohstoff, der dann in Werbung und Kontrolle zu Geld gemacht wird.

Die Plattformbetreiber produzieren nichts selbst, sondern leben allein von den Inhalten, die ihnen die Nutzer kostenlos zur Verfügung stellen. Deren Verbreitung steuern sie wiederum über intransparente Algorithmen und behalten einen Großteil der Werbeeinnahmen für sich.

Journalisten und Influencer können nur raten, mit welchen Hashtags und Reizwörtern sie besonders viele Menschen erreichen. Ob, wann und warum sie ein „Shadow-Bann“ trifft, bei dem der Algorithmus bestimmte Meldungen einfach nicht mehr in die Timeline der Nutzer spült, können sie weder nachvollziehen noch verhindern.

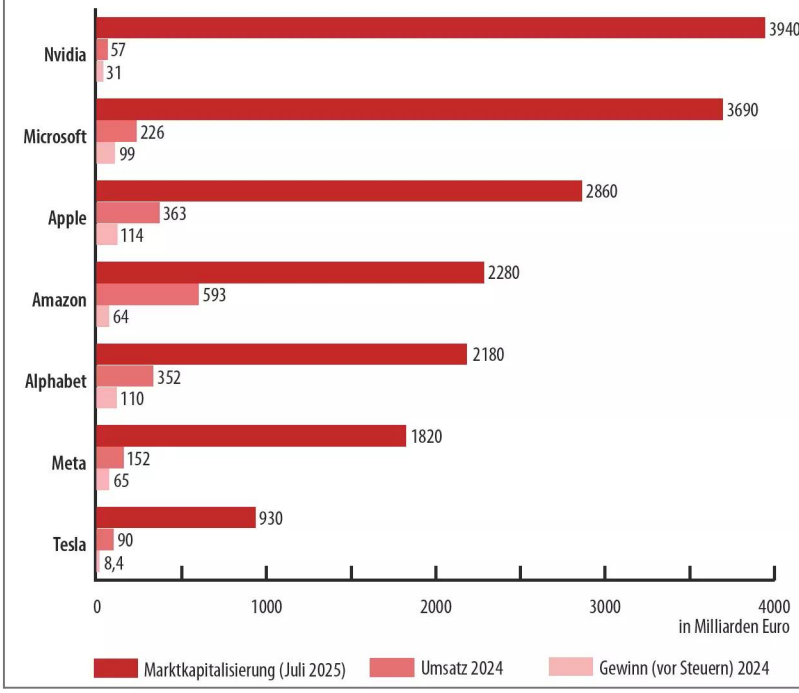
Bei Meta ist das Missverhältnis zwischen Abschöpfung und Ausschüttung besonders deutlich. Der Konzern erzielte 2024 einen Gesamtumsatz von 165 Milliarden US-Dollar, wovon 98 Prozent aus Werbung stammten. Der Gewinn belief sich im selben Jahr auf rund 62 Milliarden US-Dollar. Die Creator-Programme, mit denen Meta Influencer und Content-Produzenten entlohnt, machten im Verhältnis dazu nur einen Bruchteil aus. Ausgeschüttet wurden gerade einmal zwei Milliarden US-Dollar – weniger als zwei Prozent der Werbeeinnahmen.

Alles unter einem Dach

Mit WhatsApp verdient Meta zwar bislang kaum Geld, doch der Dienst hat eine zentrale Bedeutung für die Kontrolle über die digitale Kommunikation. Seit Jahren blockiert Meta Bemühungen, WhatsApp

Die Kapitalmacht der „glorreichen Sieben“

Die Marktkapitalisierung der Techriesen hat auch für börsennotierte Unternehmen exorbitante Höhen erklommen. Zum Vergleich: Alle 40 Unternehmen aus dem DAX sind nur etwa 2 Billionen Euro wert.



für die Interoperabilität mit anderen Messengern wie Signal oder Threema zu öffnen. Die EU hat im Rahmen des Digital Markets Act zwar vorgeschrieben, dass große Plattformen künftig Schnittstellen bereitstellen müssen, doch Meta verzögert die Umsetzung und verweist auf angebliche Sicherheitsrisiken.

Meta arbeitet seit Jahren daran, WhatsApp um Unternehmenskommunikation, Bezahlfunktionen und Handelsmöglichkeiten zu erweitern – mit dem Ziel, die App zu einer zentralen Schnittstelle für digitale Dienstleistungen auszubauen. Damit würde sich das Geschäftsmodell dem des chinesischen Vorbilds WeChat annähern, das vom Konzernimperium Tencent betrieben wird. WeChat hat in China mehr als eine Milliarde aktive Nutzer und wickelte 2023 ein Zahlungsvolumen von über 30 Billionen US-Dollar ab. Der Dienst ist nicht nur Messenger, sondern zugleich Bezahl-App, Shoppingplattform und soziales Netzwerk. Aufgrund der engen Verzahnung mit der chinesischen Regierung dient es zugleich als Werkzeug staatlicher Kontrolle über Kommunikation und Finanzflüsse.

Eine ähnliche Vision verfolgt Elon Musk mit X, dem ehemaligen Twitter. Auch hier ist das Ziel eine „Everything-App“, die Kommunikation, Medienkonsum und Zahlungsverkehr auf einer Plattform bündelt. Wie bei WhatsApp ginge es nicht nur darum, Nutzer möglichst lange im eigenen Ökosystem zu halten, sondern auch die Zahlungsströme zu kontrollieren.

Bislang fehlen X die erforderlichen Nutzerzahlen und die technische Infrastruktur, um mit WeChat vergleichbar zu sein. Doch die strategische Richtung ist die gleiche: Soziale Netzwerke und Messenger sollen sich zu allumfassenden Plattformen wandeln, die sämtliche Lebensbereiche digital erfassen und monetarisieren.

Microsofts Bürodominanz

Während Amazon, Google und Meta vor allem den Handel, die Werbung und Datenflüsse dominieren, hat Microsoft ein anderes Standbein: Es hat sich ein faktisches Monopol in der digitalen Arbeitswelt geschaffen. Mit Word, Excel, Outlook und PowerPoint verfügen die Redmonder über eine Marktdominanz, wie sie selten ein Softwarehersteller erreicht hat. In der EU laufen Microsofts Office-Programme oder deren Nachfolger in der Microsoft-365-Cloud auf 80 bis 90 Prozent aller Bürorechner. Konkurrenzprodukte wie LibreOffice fristen ein Nischendasein.

Diese Abhängigkeit erlaubt es Microsoft, die Preise fast nach Belieben anzuheben. 2022 kündigte das Unternehmen an, die Gebühren für die populären Office-365-Pakete um 10 bis 20 Prozent zu erhöhen – ein Schritt, der Analysten zufolge allein im ersten Jahr zusätzliche Einnahmen von über fünf Milliarden US-Dollar brachte. Millionen Unternehmen, Verwaltungen und Hochschulen haben keine Alternative, als zu zahlen. Die Daten liegen in Microsoft-Formaten, die gesamte Belegschaft ist auf die Programme geschult – ein Wechsel würde Milliarden kosten.

Kartellwächter in der EU und in Deutschland haben diese Entwicklung immerhin registriert. Die EU-Kommission eröffnete 2023 ein Verfahren, weil Microsoft seine Teamarbeitsplattform Teams zwangsweise an Office 365 gekoppelt hatte, was Wettbewerber wie Slack oder Zoom benachteiligte. Im Sommer 2024 sah sich Microsoft gezwungen, Teams in Europa separat anzubieten. Doch die strukturelle Marktdominanz wurde damit kaum gebrochen. Über sein Karrierenetzwerk LinkedIn hat Microsoft zudem Zugriff auf über eine Milliarde Nutzerprofile. Die Daten nutzt der Konzern, um Kommunikations- und Planungsprogramme wie Outlook und Teams enger mit den Unternehmen und dem Arbeitsmarkt zu verzahnen – ein entscheidender Vorteil gegenüber Konkurrenten.

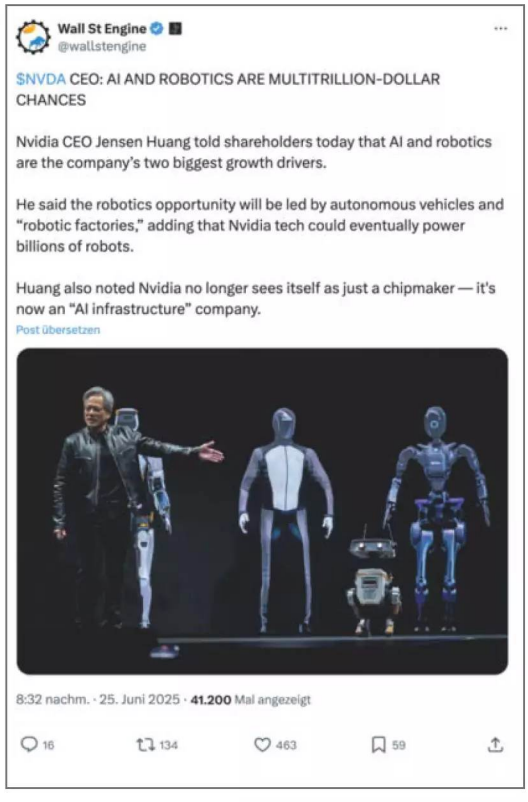
KI als Preishebel

Wie sich die Datensammelei monetarisieren lässt, zeigt Microsofts „Copilot“: Für 400 Euro pro Jahr und Nutzer erhalten Firmenkunden eine KI-Erweiterung, die Texte formuliert, E-Mails vorsortiert oder Zahlenkolonnen auswertet. In einem Unternehmen mit 10.000 Beschäftigten summieren sich die Kosten auf 4 Millionen Euro pro Jahr – zusätzlich zu den ohnehin fälligen Lizenzgebühren für Microsoft 365.

Im Unterschied dazu stellt Meta sein Sprachmodell „Llama“ kostenlos zur Verfügung. Damit will Meta von Google und Microsoft unabhängig bleiben und möglichst viele Entwickler an sein eigenes Modell binden, damit sie es etwa zur automatisierten Produktion von Werbung auf Instagram und Facebook einsetzen.

Noch macht Meta mit KI Verluste, doch es stärkt sein Kerngeschäft: den Verkauf und die gezielte Ausspielung von Werbung. Wer heute Anwendungen mit Llama baut, ist morgen auf Metas Infrastruktur angewiesen. Ob die Strategie aufgeht, ist jedoch offen: Für die riesigen Rechenzentren, die Llama erst

Bild: X



Wenn Nvidia-Chef Jensen Huang „multitrillion-dollar chances“ für KI und Robotik verspricht, ist das weniger Technologieprognose als Anlegerwerbung: Tweets wie dieser auf X funktionieren wie Hochglanzprospekte – nur billiger und mit größerer Reichweite.

ermöglichen, muss Meta jedes Jahr Dutzende Milliarden investieren. Sollte der erhoffte Nutzen ausbleiben, bliebe das Modell eine kostspielige Image-Kampagne.

Apple setzt hingegen auf Exklusivität. Unter dem Etikett „Apple Intelligence“ stattet der Konzern seine Geräte mit KI-Funktionen aus – allerdings nur die neuesten Generationen der iPhones, iPads und Macs. Der Zusatznutzen wird damit an den Kauf neuer Hardware gekoppelt. Zugleich sichert Apple seine Plattformen gegenüber Google und Meta mit dem Argument „Datenschutz“ ab. Viele Nutzer nehmen das als Vorteil wahr, doch ökonomisch erfüllt es einen anderen Zweck: Apple hält die Daten im eigenen Garten. So wird verhindert, dass Konkurrenten Zugriff auf diese Ressource erhalten und Apple-Nutzern Brücken zum Systemwechsel bauen. Den Kunden bleibt kaum etwas anderes übrig, als jedes Jahr für die teuren Geräte tiefer in die Tasche zu greifen.

Gemeinsam ist allen Konzernen, dass sie KI nicht zum Selbstzweck entwickeln, sondern als Hebel für Preiserhöhungen, Abhängigkeiten und Plattformrenten. Microsoft monetarisiert Zusatzfunktionen, Meta verschenkt Modelle, um Daten und Entwickler zu binden, und Apple koppelt KI an Premium-Geräte.

Nvidias Schaufelmonopol

Wenn Amazon im Onlinehandel systematisch Gewinne abschöpft und Microsoft zusätzliche Lizenzkosten für den KI-Einsatz im Büro erhebt, dann ist Nvidia der Lieferant, der den Goldgräbern die Schaufeln verkauft. Der Unterschied: Schaufeln gab es im 19. Jahrhundert in jedem Eisenwarenladen – die modernsten KI-Grafikprozessoren (GPUs) gibt es nur bei Nvidia.

Im Geschäftsjahr 2024 schossen die Umsätze des Konzerns von zuvor 27 auf über 60 Milliarden US-Dollar, wovon allein der Bereich Rechenzentren – also der Verkauf von KI-Chips an Amazon, Microsoft, Google und Meta – mehr als 47 Milliarden US-Dollar ausmachte. Die Bruttomarge kletterte auf über 70 Prozent – ein Wert, von dem klassische Industriekonzerne nur träumen können. Durch seine Monopolstellung kann Nvidia auch hier im ökonomischen Sinn eine Rente abschöpfen: ein Extraprofit, der allein aus der Kontrolle über einen Engpass an leistungsfähigen KI-Prozessoren entsteht.

Ein KI-Server, ausgestattet mit Nvidias Chips, hat eine Lebensdauer von drei bis fünf Jahren, bevor er technisch veraltet. Jede neue Chip-Generation – aktuell „Hopper“ (H100), bald „Blackwell“ (B100) – verdoppelt oder verdreifacht die Leistung. Für die ebenfalls als Hyperscaler agierenden Konzerne wie Amazon, Microsoft oder Google heißt das: Sie müssen ihre Rechenzentren im Rhythmus weniger Jahre erneuern, wenn sie nicht ins Hintertreffen geraten wollen. Ein Konkurrent müsste nicht nur bei der Rechenleistung der KI-Chips aufholen, sondern auch eine ähnliche Entwicklungsumgebung wie CUDA etablieren, bei der Nvidia Jahrzehnte Vorsprung hat. Auf absehbare Zeit hat Nvidia seine Rente also sicher – zumindest bis die KI-Blase platzt.

Fata Morgana der AGI

Seit Jahren überbieten sich die Tech-Konzerne mit Ankündigungen, dass schon bald eine „Artificial General Intelligence“ (AGI) entstehen werde – also eine Maschine, die klüger sei als jeder Mensch und irgendwann sogar selbstständig ganze Unternehmen steuern könnte. Für die Finanzmärkte klingt dieser Traum nach einer Maschine, die grenzenlosen Reichtum schafft, ohne dass dazu noch Arbeiter nötig wären.

In Wirklichkeit aber kann eine AGI – wenn sie überhaupt je entsteht – nur dann für die KI-Firmen profitabel sein, wenn sie in die aufgezeigten Rentenmechanismen eingebettet wird: etwa durch Abomodelle, exklusive Zugänge oder Abhängigkeiten, die Nutzer und Firmen zur Zahlung zwingen.

Doch wer soll all die immer teurer werdenden Abos bezahlen, wenn Millionen von Menschen ihre Jobs an eine AGI verlieren? Arbeitslose und prekär Beschäftigte können sich keine teuren Smartphones und KI-Abos leisten. Damit bricht den Konzernen die Nachfrage weg. Ohne Einnahmen fehlen wiederum die Mittel für neue Server und Modelle.

Die AGI-Versprechen funktionieren wie eine Fata Morgana: Sie sollen die Erwartungshaltung der Börse hoch halten und weiterhin Milliardeninvestitionen rechtfertigen. Denn ohne den Traum von der AGI fehlt den Investoren der Glaube, dass sich ihre Wetten auf die Zukunft irgendwann einmal auszahlen könnten.

Platzt die Blase?

Nvidias Börsenwert überschritt im Sommer 2025 die Marke von 4 Billionen US-Dollar. Doch die Erwartungen der Investoren haben sich hier augenscheinlich von den realen Gewinnen entkoppelt, die 2024 bei rund 34 Milliarden Dollar (31 Milliarden

Euro, vor Steuern) lagen. Die Situation erinnert an den Eisenbahnboom zu Beginn des industriellen Zeitalters. Mitte des 19. Jahrhunderts schien die Eisenbahn das Allheilmittel für Wirtschaft und Gesellschaft zu sein. In Großbritannien schossen Aktiengesellschaften wie Pilze aus dem Boden, die verkündeten, Schienen bis in die entlegensten Täler zu bauen.

Jede neue Bahnlinie versprach Zugang zu Märkten, schnellere Warenströme, steigende Grundstückspreise entlang der Strecke. Anleger pumpten jedes Jahr Summen in Höhe von 15 bis 20 Prozent des britischen Bruttoinlandsprodukts in den Schienenausbau. Doch bald zeigte sich, dass viele Linien überflüssig waren oder nie rentabel betrieben werden konnten. 1847 platzte die Blase, Aktienkurse brachen ein, viele Gesellschaften gingen bankrott.

Heute erinnert die KI-Euphorie stark an diese Episode. Amazon, Google, Microsoft oder Meta investieren Dutzende Milliarden pro Jahr in Rechenzentren, die so viel Strom verbrauchen wie Kleinstädte. Doch schon jetzt berichten Unternehmen, dass die erhofften Produktivitätsschübe ausbleiben, während die Cloud-Rechnungen Monat für Monat steigen. Jeder Konzern verkündet, dass seine KI-Killer-App die Geschäftswelt revolutioniert. Solche Rhetorik ist typisch für Blasen – und endete in der Geschichte fast immer in einem Crash.

Dieser wäre aber kein reinigendes Gewitter. Die M7 kämen wohl kaum ins Wanken, denn sie sind heute wesentlich kapitalstärker als die Gesellschaften im 19. Jahrhundert. Ein Crash, der viele kleinere und mittlere Unternehmen in den Abgrund reißt, würde ihre Macht wahrscheinlich sogar stärken – vor allem, wenn die Zentralbanken als Antwort auf eine neue Wirtschaftskrise erneut frische Milliarden in die Märkte pumpen, die am Ende wieder bei den großen Sieben landen.

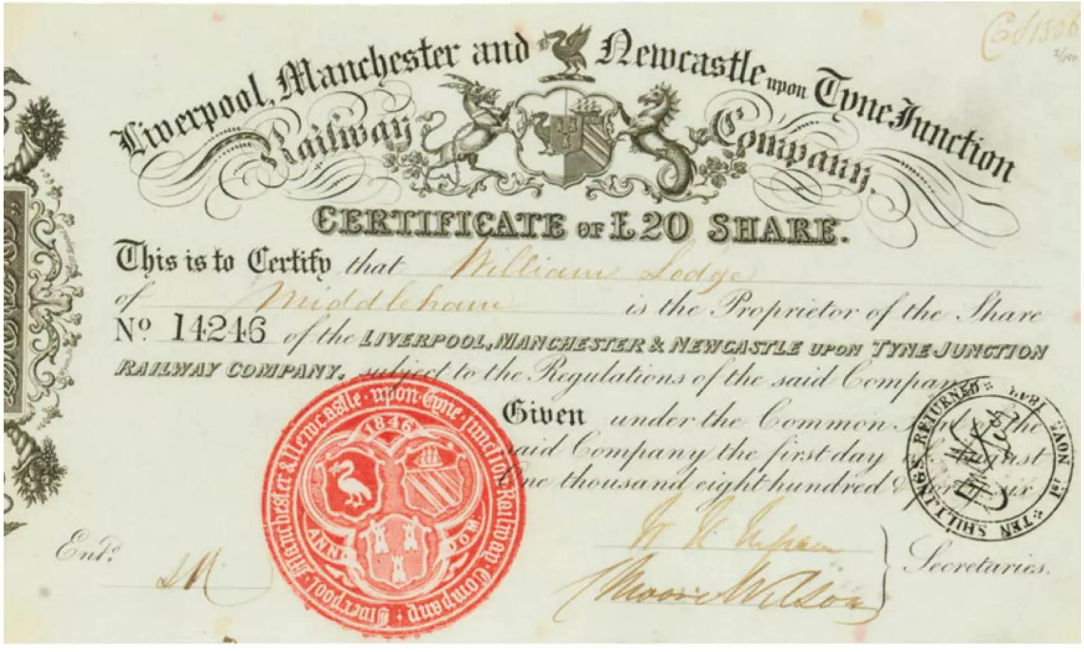


Bild: EDHAC, Public Domain

Wie KI heute lockten damals Eisenbahnen Investoren: 1846 setzten Anleger große Hoffnungen auf den Ausbau der Streckennetze und zeichneten massenhaft Aktien neuer Gesellschaften – auch Anteilsscheine der Liverpool, Manchester & Newcastle-upon-Tyne Junction Railway Co. Nach dem Platzen der Blase wurde die Gesellschaft nur zwei Jahre später bereits wieder aufgelöst.

Gesellschaft mit beschränkter Hoffnung

Die Situation erscheint also ziemlich düster, denn Alternativen können unter diesen Umständen nur schwer keimen. Unter der Dominanz der M7 leiden nicht nur deren Beschäftigte und Millionen ausgelagerter Clickworker im globalen Süden, die für Hungerlöhne Daten für KI-Modelle bereinigen (mehr dazu ab Seite 168 und in [8]). Auch die Milliarden Nutzerinnen und Nutzer der Plattformen tragen mit ihren Inhalten und Daten tagtäglich zum Wert der Konzerne bei, ohne daran beteiligt zu werden.

Im Unterschied zu klassischen Arbeitskämpfern stehen sie dabei nicht allein. Sie sitzen im selben Boot wie mittelständische Unternehmen und Händler, die unter den steigenden Abgaben an Amazon, Google oder Microsoft ächzen und nach Wegen suchen, ihre Abhängigkeit zu reduzieren. Daraus könnten sich neue Allianzen bilden: zwischen Angestellten, Konsumenten und Unternehmern, die ein gemeinsames Interesse daran haben, den M7 etwas entgegenzusetzen.

Verbraucher haben bereits gezeigt, dass kollektive Aktionen Wirkung entfalten können. Kampagnen wie „Make Amazon Pay“ (makeamazonpay.com), koordiniert von der UNI Global Union (uniglobalunion.org/de), legten am Black Friday weltweit Logistikzentren lahm und verursachten empfindliche Umsatzeinbußen. Solche Aktionen gewinnen an Schlagkraft, wenn sie von kleineren und mittleren Unternehmen flankiert werden, die politischen Druck auf Kartellwächter und Gesetzgeber ausüben. Aufsichtsbehörden müssten jedoch eingreifen können, sobald sich eine marktbeherrschende Stellung abzeichnet – und nicht erst, wenn unlautere Methoden nachweisbar sind.

Langfristig reicht eine bloße Regulierung, wie sie etwa Shoshana Zuboff fordert, jedoch nicht aus – sie lindert die Symptome, packt aber nicht deren Ursachen an. Nötig ist vielmehr eine Neuorientierung der Finanz- und Wirtschaftspolitik. Neoliberale Pläne,

Literatur

[1] Yanis Varoufakis, Techno-Feudalismus, Was den Kapitalismus tötete, Kunstmann 2024

[2] Cédric Durand, How Silicon Valley Unleashed Techno-feudalism, Verso 2024

[3] Adam Smith, Der Wohlstand der Nationen, Untersuchung über das Wesen und die Ursachen des Volkswohlstandes, Zweitausendeins 2009

[4] Thomas Piketty, Das Kapital im 21. Jahrhundert, C.H.Beck 2016

[5] Nick Srnicek, Plattform-Kapitalismus, Hamburger Edition, 2018

[6] Phillip Staab, Digitaler Kapitalismus, Markt und Herrschaft in der Ökonomie der Unknappheit, Suhrkamp 2019

[7] Shoshana Zuboff, Das Zeitalter des Überwachungskapitalismus, Campus 2025

[8] Ingo Dachwitz, Sven Hilbig, Digitaler Kolonialismus, Wie Tech-Konzerne und Großmächte die Welt unter sich aufteilen, C.H.Beck 2025

[9] Theodor W. Adorno, Minima Moralia, Reflexionen aus dem beschädigten Leben, Suhrkamp 2003

Europa durch Deregulierung und steuerfinanzierte „Unicorn-Zucht“ wettbewerbsfähig zu machen, verkennen die Realität: Zalando ist keine Alternative zu Amazon, und in Paris oder Berlin mit Steuermilliarden hochgepäppelte Start-ups enden meist als Übernahmekandidaten der M7 oder wandern in die USA ab.

Stattdessen müsste die öffentliche Unterstützung auf Modelle zielen, die sich dem Zugriff des globalen Kapitals entziehen. Varoufakis schlägt unter anderem Genossenschaften nach dem Prinzip „ein Mitglied, eine Stimme“ vor, wie es die International Cooperative Alliance (ICA, ica.coop) propagiert. Bei ihnen entscheidet nicht die Menge an Anteilsscheinen über Macht und Einfluss, sondern jedes Mitglied hat eine gleichwertige Stimme. Beispiele wie die spanische Mondragón-Kooperative oder digitale Plattformgenossenschaften wie Fairmondo (siehe ct.de/wuhg) zeigen, wie es funktionieren könnte. So ließe sich zumindest verhindern, dass staatliche Milliarden am Ende wieder in den Taschen der M7 landen.

Kritiker weisen jedoch darauf hin, dass sich Löhne und Arbeitsbedingungen in solchen Kooperativen kaum bessern, solange sie im Wettbewerb mit kapitalstarken Konzernen stehen. Ein Dilemma, das Theodor W. Adorno in seiner „Minima Moralia“ auf den Punkt brachte: „Es gibt kein richtiges Leben im falschen.“ [9]

Der britische Ökonom Nick Srnicek sieht deshalb nur einen Ausweg: Die Kontrolle über die Plattformen muss von den Privatfirmen auf öffentliche oder genossenschaftliche Modelle übertragen werden. Der französische Ökonom Cédric Durand teilt diese Einschätzung, warnt jedoch vor einer simplen Verstaatlichung, weil sie digitale Macht in ein staatliches Monopol verschieben würde – mit der Gefahr eines autoritären Überwachungsstaats.

Eine Blaupause für eine emanzipatorische Neuordnung der digitalen Plattformen fehlt bislang. Ob sich zumindest einzelne Ansätze verwirklichen können, hängt davon ab, ob es gelingt, die unterschiedlichen Akteure – Beschäftigte, Konsumenten, Mittelstand – mit ihren gemeinsamen Interessen zu bündeln. Nur gemeinsamer Druck kann die Machtkonzentration der M7 aufbrechen. (hag) ct

Harvard-Studie und weitere Links:
ct.de/wuhg

Bild: Martina Bruns/heise medien

Wie Big-Tech Mensch und Umwelt ausnutzt

Im Wettlauf um das beste KI-Modell überbieten sich die großen Digitalunternehmen wie Meta, Amazon und Microsoft mit Investitionen. Die Folge: Sie brauchen immer mehr Energie, Wasser und Ressourcen, der Klimaschutz bleibt auf der Strecke.

Von **Greta Friedrich**

Ein paar US-Unternehmen dominieren weite Teile der globalen digitalen Infrastruktur. Die „Magnificent Seven“ (M7) sind nicht nur Marktbeherrscher und Datensammler, sondern belasten zudem Umwelt und Menschen durch ihren immensen Ressourcenverbrauch. Doch die Lage ist unübersichtlich, denn Lieferketten sind komplex, Herstellerangaben oft intransparent und Standards häufig schlicht nicht vorhanden. So können beispielsweise Verbraucher die tatsächlichen Kosten einer KI-Abfrage nicht abschätzen.

Klar ist: Nvidia fährt seit Jahren Umsatzrekorde ein, vor allem mit Produkten für Rechenzentren wie zum Beispiel den mittlerweile unverzichtbaren KI-Beschleunigern. Die Produktion der dafür benötigten Halbleiter verbraucht viele Rohstoffe wie Silizium, seltene Erden oder Wasser und stößt große Mengen Treibhausgase aus [1]. Die übrigen Sechs der M7 (Microsoft, Apple, Amazon, Alphabet, Meta und Elon Musks Firmenkonglomerat) nutzen solche Chips für ihre KI-Anwendungen und -Services. Sie wetteifern mit anderen Anbietern wie OpenAI und Deepseek

darum, wer die beste, größte und beliebteste KI entwickelt – und überbieten einander mit Investitionen, Bauprojekten und Rechenleistung. Welche Folgen diese Gigantomanie der großen Tech-Firmen für die Natur, das Klima und Millionen Menschen hat, umreißt dieser Artikel.

Der Markt boomt

In den kommenden Jahren wird sehr viel Geld in den Aufbau neuer Rechenzentren fließen. Eine Anfang August 2025 vom Beratungsunternehmen McKinsey & Company veröffentlichte Analyse schätzt, dass Unternehmen bis zum Jahr 2030 weltweit fast 7 Billionen US-Dollar in die Infrastruktur von Rechenzentren pumpen werden. Die Boston Consulting Group sieht vor allem Unternehmen wie Meta, Microsoft, Google und Amazon als Wachstumstreiber: Hyperscaler wie sie würden von 2023 bis 2028 etwa 60 Prozent des Wachstums der Rechenzentrumsbranche generieren.

In den vergangenen Monaten verkündeten die großen Digitalunternehmen immer neue Milliardeninvestitionen in neue (KI-)Rechenzentren. Meta-Chef Mark Zuckerberg beispielsweise spart bei seinen Plänen nicht mit Superlativen: Anfang 2025 kündigte er ein neues Rechenzentrum an, dessen Fläche halb so groß werde wie Manhattan; im Sommer legte er nach und avisierte weitere riesige Zentren, benannt nach Titanen aus der griechischen Mythologie.

In diesem Wachstum sieht das Beratungsunternehmen McKinsey zwar eine große wirtschaftliche Chance für die geplanten Standorte von Rechenzentren. Seine Analyse betont aber auch Risiken für die Standorte, wie eine deutlich stärkere Belastung des Stromnetzes, Wasserknappheit und die Gefahr, ihre Nachhaltigkeitsziele zu verfehlen.

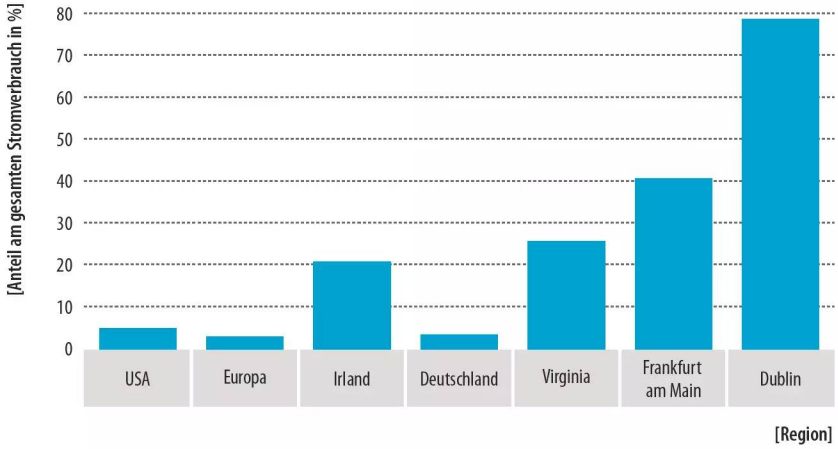
Stromnetze ächzen

In manchen Regionen ballen sich Rechenzentren bereits jetzt derart, dass sie für einen beträchtlichen Teil des dortigen Stromverbrauchs verantwortlich sind. Das Öko-Institut, eine private Forschungseinrichtung mit Fokus auf Nachhaltigkeitsthemen und Sitz in Freiburg im Breisgau, fertigte 2025 im Auftrag von Greenpeace eine Metastudie zu den Umweltauswirkungen künstlicher Intelligenz an. Diese wertet unter anderem Analysen von McKinsey und der Internationalen Energieagentur (IEA) aus: In den USA hatten Rechenzentren demnach schon 2023 geschätzt einen Anteil von über 5 Prozent am gesamten Stromverbrauch.

Bild: Öko-Institut/heise medien

Viele Rechenzentren, hoher Stromverbrauch

Anteil des Stromverbrauchs von Rechenzentren am gesamten Stromverbrauch in der jeweiligen Region im Jahr 2023 (geschätzt).



Im US-Bundesstaat Virginia, wo vor allem im Norden besonders viele Rechenzentren angesiedelt sind, liegt die Quote bei 26 Prozent. Einige europäische Städte übertreffen diese noch: In London und Frankfurt am Main liegt die Quote bei um die 40 Prozent, in Dublin sogar bei fast 80 Prozent. Diese Anteile seien allerdings mit einer gewissen Vorsicht zu behandeln, da für den Stromverbrauch von Rechenzentren und den der jeweiligen Regionen unterschiedliche Quellen herangezogen wurden, so das Öko-Institut.

Lokale Stromnetze sind für die steigende Belastung nicht immer gerüstet. Irland reguliert daher seit 2021 den Bau neuer Rechenzentren: Bevor sie ans Netz dürfen, sollen die lokalen Netzbetreiber zum Beispiel prüfen, wie belastet das Stromnetz in der betreffenden Region bereits ist und ob der Rechenzentrumsbetreiber seinen Verbrauch auf Anfrage drosseln könnte.

Um ihren steigenden Energiebedarf zu stillen und dennoch ihre Nachhaltigkeitsziele zu erreichen, nutzen Big-Tech-Unternehmen unterschiedliche Strategien. Microsoft sichert sich beispielsweise Strom aus nachhaltigen Quellen über Power-Purchase-Agreements mit den Erzeugern und finanziert ein altes Atomkraftwerk in den USA, Meta sichert sich dort für 20 Jahre die gesamte Energieproduktion eines Kernkraftwerks und Google lässt gleich drei neue Reaktoren bauen. Außerdem investieren sie in Zukunftstechniken, Google zum Beispiel in Kernfusion und sogenannte Small Modular Reactors (SMR). In die Entwicklung der kleinen Atomkraftwerke investiert auch Amazon, doch bisher gibt es weder marktreife SMR noch funktionierende Kernfusion.

Engpässe drohen

Ende 2024 warnte die North American Electric Reliability Corporation (NERC), dass in Nordamerika Energieengpässe drohen, bedingt durch den Bau neuer energieintensiver Rechenzentren für künstliche Intelligenz und Krypto-Mining. Die NERC ist eine gemeinnützige Organisation, die unter staatlicher Aufsicht die Stromnetze koordiniert. Sie mahnte an, die Kapazitäten für Energieerzeugung und -übertragung dringend auszubauen.

Die Kosten für diesen Ausbau könnten Stromerzeuger und Netzbetreiber auf kleinere Unternehmen und Verbraucher umlegen, befürchtet die New York Times in einem aktuellen Bericht. Es sei denn, die US-Regierung nähme die Big-Tech-Unternehmen in die Pflicht, die mit ihren neuen Rechenzentren und Kraftwerken das Stromnetz belasten.

Anfang 2025 deutete eine Untersuchung des Finanzdienstes Bloomberg in den USA darauf hin, dass es im Umfeld von KI-Rechenzentren verstärkt zu sogenannten Oberschwingungen im Stromnetz kommen könnte, die wiederum die Stromqualität beeinträchtigen und zu Ausfällen führen können. Netzbetreiber und einzelne Stromversorger zweifelten die Ergebnisse dieser Analyse allerdings an.

KI treibt den Energiebedarf …

In den nächsten Jahren wird der weltweite Strombedarf von Rechenzentren drastisch ansteigen, prognostiziert unter anderem die IEA (siehe Bild unten). Ein Grund dafür ist die fortschreitende Verbreitung

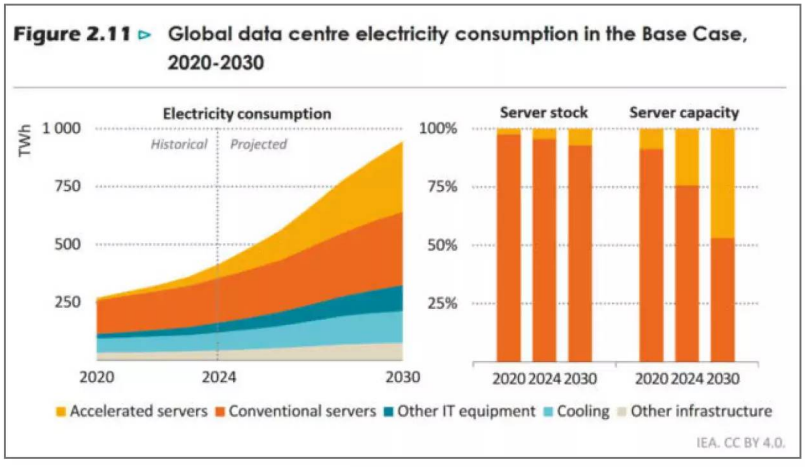
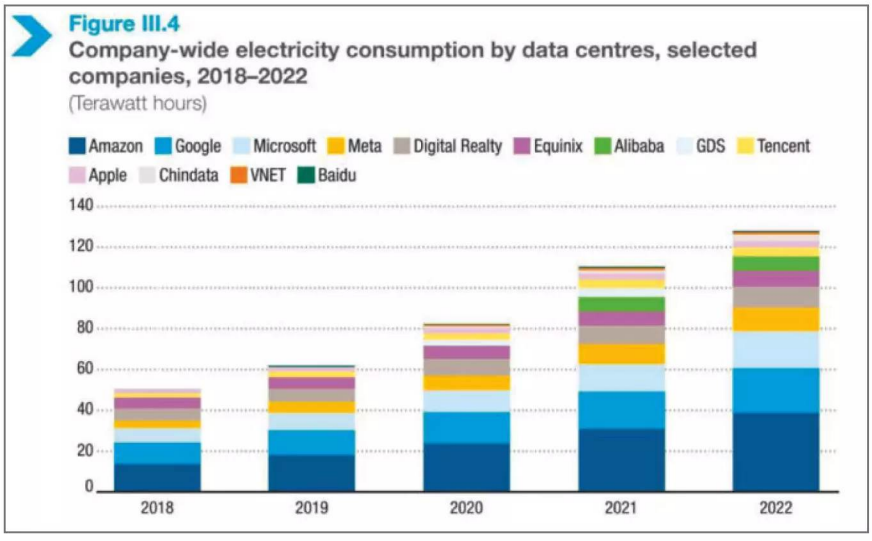


Bild: IEA

Innerhalb von zehn Jahren könnte sich der weltweite Energiebedarf von Rechenzentren mehr als verdreifachen. Den größten Anteil daran haben beschleunigte Server, obwohl sie nur einen kleinen Teil des Serverbestands ausmachen. KI-Anwendungen laufen meist auf diesen Servern.

Rechenzentren von Amazon, Google, Microsoft und Meta benötigten 2022 mehr als doppelt so viel Energie wie noch 2018 – und da begann erst der offene Wettlauf um die beste KI.

Bild: UNCTAD



von KI-Anwendungen, insbesondere von generativer KI. Multimodale Modelle sind, anders als kleine, spezialisierte Modelle, für verschiedene Zwecke nutzbar. Sie können zum Beispiel sowohl Text als auch Bilder und Videos erzeugen, brauchen aber deutlich mehr Energie für das Training und den Betrieb.

Für solch komplexe Berechnungen sind spezielle Chips nötig, die sogenannten KI-Beschleuniger. Derzeit ist zum Beispiel der Blackwell-Chip von Nvidia besonders begehrt. Zwar werden KI-Beschleuniger mit jeder Generation effizienter. Doch die KI-Modelle wachsen schneller und die konkurrierenden KI-Anbieter wollen immer größere Modelle immer schneller trainieren als die Konkurrenten.

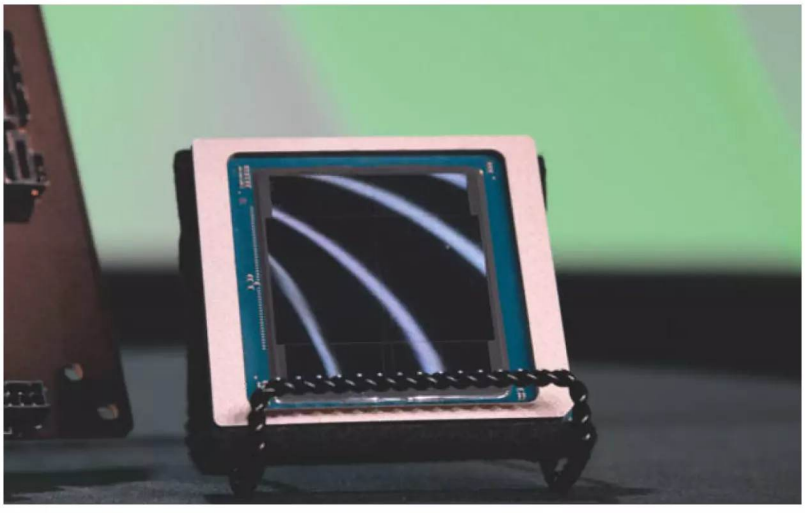
Daher bauen sie eilig immer gewaltigere Rechenzentren auf, der Energiebedarf steigt. Zum Vergleich: In Deutschland haben die meisten Rechenzentren derzeit eine Anschlussleistung für Server von weniger als 5 Megawatt. In den USA werden hingegen bereits viele Rechenzentren mit Anschlussleistungen von jeweils über 1 Gigawatt gebaut. Für 2030 schätzen Bitkom und das Borderstep Institut, dass die Summe der Anschlussleistungen aller deutschen Rechenzentren bei 4,8 Gigawatt liegen wird. In den USA sollen es dann bereits 95 und in China 64 Gigawatt sein.

… und den Wasserverbrauch

Auch in der Chipherstellung [1] ist sehr viel Energie nötig: Laser für das photolithografische Verfahren, Vakuumpumpen für mehrere Fertigungsschritte sowie Lüftung und Filter für die Reinräume brauchen viel Strom. Wasser ist ebenfalls wichtig für die Halbleiterherstellung, es dient einerseits zur Prozesskühlung und andererseits zur Reinigung. Mittlerweile gehen Hersteller dazu über, Wasser zu sparen, indem sie es aufbereiten und wiederverwenden.

Auch viele Rechenzentren benötigen Wasser: Weil die Chips aktueller KI-Systeme extrem dicht gepackt in Serverracks gestapelt liegen, lassen diese sich praktisch nur noch mit Wasser kühlen statt mit Luft. Dank Wasserkühlung sind Serverracks mit aktuell 145 Kilowatt und bald sogar 1 Megawatt Leistung möglich – früher waren 3 bis 8 Kilowatt üblich. So kann ein einzelnes Rechenzentrum enorme absolute Leistungen erreichen, allerdings steigt der Wasserverbrauch für die Rückkühlung des intern zirkulierenden Kühlwassers durch Verdunstung.

Dadurch, dass infolge des KI-Booms immer mehr und immer größere Rechenzentren entstehen, steigt der Wasserverbrauch umso stärker: Das Öko-Institut prognostiziert, dass sich der globale Wasserverbrauch von Rechenzentren von 2023 bis 2030 mehr



Halbleiterhersteller benötigen viel Energie und Wasser. Außerdem verarbeiten sie mehr als die Hälfte aller chemischen Elemente, die das Periodensystem aufführt. Das Bild zeigt den Blackwell-Chip von Nvidia.

Nvidia stellt die derzeit begehrtesten KI-Beschleuniger her. Der abgebildete Rackeinschub (Bull Sequana XH3515) ist mit acht GH200 Grace Hopper Superchips bestückt. Auf der linken Seite wurden die Kühlblöcke demontiert, sodass vier CPUs (hellgrau) und vier GH200 (dunkelgrau) erkennbar sind. Die rechte Hälfte ist vollständig montiert, inklusive Flüssigkühlung.

als verdreifachen könnte – der Verbrauch in Zentren, die sich auf KI-Anwendungen spezialisiert haben, könnte sich sogar verzehnfachen.

Lithium kostet Wasser

Doch nicht nur Rechenzentren und Chipeinsatz der großen Tech-Firmen erfordern einen hohen Ressourcenverbrauch, sondern auch ihre Elektronikgeräte für Endnutzer. Das Leichtmetall Lithium ist essenziell für Lithium-Ionen-Akkus, die zum Beispiel in den Smartphones von Apple oder den Autos von Tesla stecken. Besonders häufig kommt das Material im sogenannten Lithium-Dreieck in Südamerika vor, an den Grenzen von Chile, Bolivien und Argentinien.

Um Lithium zu gewinnen, pumpt man dort Sole aus dem Erdboden in große flache Becken, aus denen das meiste Wasser verdunstet. Aus der konzentrierten Sole wird dann in einem chemischen Prozess pulverförmiges Lithiumcarbonat gewonnen. Brot für die Welt veröffentlichte 2018 einen Report über die Lithiumförderung: Demnach verdunsten zur Herstellung von einer Tonne Lithium circa zwei Millionen Liter Wasser.

Zwar eigne sich die Sole selbst nicht zum Trinken oder zum Bewässern von Pflanzen. Durch die starke Entnahme von Wasser sinke aber der Grundwasserspiegel insgesamt, so zum Beispiel in der Gegend um den Atacama-Salzsee in Chile. Während das Wasser knapper werde, steige der Salzgehalt in der Landschaft und das Ökosystem verändere sich. All

Das intern zirkulierende Kühlwasser strömt durch Kühltürme auf dem Dach des Rechenzentrums oder daneben. Reicht die Luft nicht zur Rückkühlung, werden die Kühlerlamellen mit Wasser besprüht, das verdampft und dadurch kühlt.

das schade den Menschen, Tieren und Pflanzen, die dort leben, so der Report.

Datenarbeit mit psychischen Folgen

Das ungebremste Wachstum der Tech-Firmen schadet auch Menschen, die für sie arbeiten. Der KI-Boom beruht in hohem Maße auf kleinteiliger Datenarbeit, die in Hochglanzankündigungen neuer Sprachmodelle oder Agenten unsichtbar bleibt. Im Buch „Digitaler Kolonialismus. Wie Tech-Konzerne und Großmächte die Welt unter sich aufteilen“ [2] nutzen die Autoren daher den Begriff „Geisterarbeit“ (der wiederum auf ein gleichnamiges Buch zurückgeht).

Ingo Dachwitz und Sven Hilbig haben für ihr Buch mit Betroffenen gesprochen. Diese haben verschiedene Arbeiten für die Tech-Firmen verrichtet: zum Beispiel Trainingsdaten für ChatGPT klassifiziert und dabei auch verstörende Inhalte gelabelt, Metadaten zu Bildern fürs Training selbstfahrender Autos ergänzt oder belastende Social-Media-Inhalte für Meta moderiert. Angestellt waren die Datenarbeiter bei einem Outsourcing-Unternehmen in Kenia.

Die Betroffenen berichten im Buch von extremer psychischer Belastung, weil sie sich bei der Arbeit beispielsweise unzählige Hasspostings, Pornografie in Text und Bild oder Videos von Folter und Hinrichtungen ansehen mussten. Mit Angehörigen darüber reden durften sie nicht und bekamen auch von ihren Arbeitgebern kaum psychologische Unterstützung. Stattdessen waren sie hohem Zeit- und Leistungsdruck ausgesetzt; ein Content-Moderator für Meta musste zum Beispiel 600 bis 700 Fälle pro Tag beurteilen. Das alles für unter zwei Dollar pro Stunde. Das habe System, so Dachwitz und Hilbig in ihrem Buch. Es gebe bei allen großen Tech-Outsourcing-Unternehmen Berichte über Ausbeutung.

Microwork by Amazon

Ein anderes Arbeitsmodell seien Microwork-Plattformen, die Kleinstaufträge an freie Arbeitnehmer vermitteln. Diese arbeiten selbstständig, in der Regel am heimischen Rechner, und bewerten zum Beispiel Restaurants oder annotieren Daten fürs KI-Training. Pro Aufgabe bekämen die Arbeiter in der Regel weniger als einen Cent. Sie seien „die Tagelöhner:innen des Digitalzeitalters“, stellen die Autoren des Buchs fest. Amazon habe dieses Arbeitsmodell Anfang der 2000er-Jahre etabliert, um Dubletten aus seinem Marktplatz zu fischen.

Die Buchautoren sprachen mit der Entwicklungsökonomin Uma Rani, die bei der Internationalen Arbeitsorganisation der UN (ILO) zu digitaler Arbeit und Outsourcing forscht. Nach ihrer Einschätzung müsse man von zig Millionen Geisterarbeitern in der Tech-Industrie ausgehen. Das Ausmaß sei aber sehr schwierig zu beziffern, da es kaum offizielle Zahlen dazu gebe.

Mehr Transparenz schaffen

Die großen Digitalunternehmen sind dafür mitverantwortlich, dass immer mehr Energie, Wasser sowie Rohstoffe benötigt, Flächen versiegelt sowie Elektroabfälle produziert werden, dass die Energiewende stockt und Millionen Menschen unter Ausbeutung leiden. Doch es gibt kaum global geltende Regeln, an die sich die Konzerne halten müssten. Es fehlen Standards dafür, wie sie Energieaufwand und Wasserverbrauch, die Herkunft und Entsorgung von Ressourcen sowie Arbeitsbedingungen bei outgesourcten Tätigkeiten transparent anzugeben haben.

Deshalb mogeln sich Techkonzerne ihre Zahlen zurecht, lassen Angaben weg oder hübschen sie auf. So hat Google beispielsweise kürzlich in einer Studie dargelegt, dass ein Gemini-Prompt lediglich eine

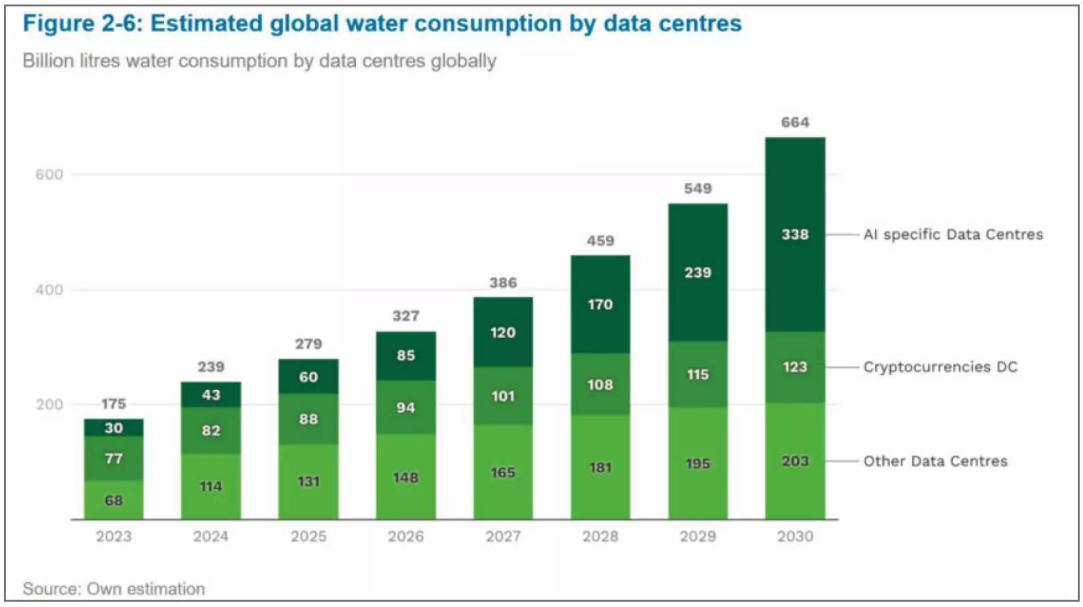


Bild: Öko-Institut

Da immer mehr Rechenzentren für KI-Anwendungen gebaut werden, steigt deren Gesamtwasserverbrauch stark.

sehr geringe Menge Wasser verbrauche und wenig Treibhausgase ausstoße. Forscher kritisierten daraufhin Ende August, dass Google entscheidende Daten einfach nicht berücksichtigt habe.

Es fehle außerdem die Angabe, wie viele Prompts Gemini pro Tag bearbeitet – das wäre wichtig zu wissen, um seinen Ressourcenverbrauch wirklich einzuschätzen. Ein einzelner Prompt fällt nicht ins Gewicht, ähnlich wie eine einzelne, dünne Plastiktüte am Obstregal im Supermarkt. Sie verursacht keine größeren Probleme, die Gesamtmenge an Verpackungsmüll aber sehr wohl.

Ein anderes Beispiel ist die Apple Watch: Ebenfalls Ende August gewann die Deutsche Umwelthilfe (DUH) ein Gerichtsverfahren gegen Apple. Das Landgericht Frankfurt entschied, dass Apple seine Uhren hierzulande künftig nicht mehr als CO_2-neutral bewerben dürfe. Denn die Uhr sei nicht klimaneutral produziert, sondern Apple nutze Kompensationsprojekte. Dort werde CO_2 zwar gebunden, allerdings nur für wenige Jahre in kommerziellen Eukalyptusplantagen, die der Betreiber später wieder abholzt.

KI: Problem oder Lösung?

Angesichts ihres massiven Wachstums ist es schwer vorstellbar, dass die großen Digitalunternehmen in absehbarer Zeit tatsächlich klimagerecht oder gar nachhaltig wirtschaften könnten. Im Juni stellte der Report „Greening Digital Companies“ von der Internationalen Fernmeldeunion (ITU) und der gemeinnützigen World Benchmarking Alliance (WBA) fest, dass zum Beispiel bei Amazon, Microsoft, Alphabet und Meta die betrieblichen Emissionen seit 2020 um durchschnittlich 150 Prozent gestiegen sind. Die vier gehörten außerdem zu den zehn energiehungrigsten der betrachteten Unternehmen (200 führende Unternehmen im Sektor der Informations- und Kommunikationstechnik).

Die bereits zitierte IEA-Untersuchung zu Energie und KI beschäftigt sich auch mit der Frage, wie KI den weltweiten Energiehunger tatsächlich steigert. Einerseits hoffe man darauf, dass die selbstlernenden Systeme dabei helfen können, Prozesse zu optimieren und Emissionen zu reduzieren. Andererseits seien auch Rebound-Effekte möglich: Effizienzsteigerungen würden dadurch zunichtegemacht, dass sie zu mehr Konsum führen.

Ein solcher Effekt lässt sich bereits an der Basis bei der Modellentwicklung beobachten: Das Potenzial der immer effizienteren Chiptechnik nutzen die Digitalunternehmen nicht, um ihren Energiebedarf zu senken. Sondern dazu, das nächste KI-Modell mit noch mehr Parametern auszustatten und mit noch mehr Daten zu trainieren. (gref) *ct*

Literatur

[1] Christof Windeck, Ökochips, Wie die hohen Umweltlasten der Halbleiterfertigung sinken sollen, c't 26/2023, S. 126

[2] Ingo Dachwitz und Sven Hilbig, Digitaler Kolonialismus, Wie Tech-Konzerne und Großmächte die Welt unter sich aufteilen, ISBN: 3406823025, S. 19 ff.

Verwendete Quellen und weitere Informationen:

ct.de/ws88

IMPRESSUM

Redaktion

Postfach 61 04 07, 30604 Hannover
Karl-Wiechert-Allee 10, 30625 Hannover
Telefon: 05 11/53 52-300
Telefax: 05 11/53 52-417
Internet: www.heise.de

Leserbriefe und Fragen zum Heft:
sonderhefte@ct.de

Die E-Mail-Adressen der Redakteure haben die Form xx@heise.de oder xxx@heise.de. Setzen Sie statt „xx“ oder „xxx“ bitte das Redakteurs-Kürzel ein. Die Kürzel finden Sie am Ende der Artikel und hier im Impressum.

Chefredakteur: Torsten Beeck (tbe, verantwortlich für den Textteil), Dr. Volker Zota (vza)

Konzeption: Hartmut Gieselmann (hag)

Koordination: Jobst Kehrhahn (keh, Leitung), Pia Groß (piae)

Redaktion: Jo Bager (jo), Holger Bleich (hob), Ronald Eikenberg (rei), Greta Friedrich (gref), Hartmut Gieselmann (hag), Arne Grävelmeyer (agr), Nico Jurran (nij), André Kramer (akr), Jan Mahn, (jam), Dr. Sabrina Patsch (spa), Sylvester Tremmel (syt), Andrea Trinkwalder (atr)

Mitarbeiter dieser Ausgabe: Rebecca Haar, Tanja Kunesch, Kai Schwirzke, Sebastian Springer

Assistenz: Susanne Cölle (suc), Tim Rittmeier (tir), Martin Triadan (mat)

DTP-Produktion: Vanessa Bahr, Dörte Bluhm, Lara Bögner, Beatrix Dedek, Madlen Grunert, Emilie Hertzke, Cathrin Kapell, Steffi Martens, Lisa Reich, Marei Stade, Matthias Timm, Christiane Tümmler, Nicole Wesche

Digitale Produktion: Christine Kreye (Leitung), Thomas Kaltschmidt, Martin Kreft, Pascal Wissner

Illustration, Fotografie: Thorsten Hübner, Albert Hulm, Moritz Reichartz

Titel: Steffi Martens, www.freepik.com

Verlag

Heise Medien GmbH & Co. KG
Postfach 61 04 07, 30604 Hannover
Karl-Wiechert-Allee 10, 30625 Hannover
Telefon: 05 11/53 52-0
Telefax: 05 11/53 52-129
Internet: www.heise.de

Herausgeber: Christian Heise, Ansgar Heise, Christian Persson

Geschäftsführer: Ansgar Heise, Beate Gerold

Mitglieder der Geschäftsleitung: Jörg Mühle, Falko Ossmann

Anzeigenleitung: Michael Hanke (-167)
(verantwortlich für den Anzeigenteil),
www.heise.de/mediadaten/ct

Anzeigenverkauf: Verlagsbüro ID GmbH & Co. KG,
Tel.: 05 11/61 65 95-0, www.verlagsbuero-id.de

Leiter Vertrieb und Marketing: André Lux (-299)

Service Sonderdrucke: Julia Conrades (-156)

Druck: Firmengruppe APPL Druck GmbH & Co. KG,
Senefelder Str. 3–11, 86650 Wemding

Vertrieb Einzelverkauf:
DMV DER MEDIENVERTRIEB GmbH & Co. KG
Meßberg 1
20086 Hamburg
Tel.: 040/3019 1800, Fax: 040/3019 145 1815
E-Mail: info@dermedienvertrieb.de
Internet: dermedienvertrieb.de

Einzelpreis: € 14,90; Schweiz CHF 27.90;
Österreich € 16,40; Luxemburg € 17,10

Erstverkaufstag: 24.10.2025

Bild: Martina Bruns/KI/heise medien

Wie Europa unabhängiger wird

Sich kampflos dem Schicksal der digitalen Abhängigkeit zu ergeben, wäre fatal. Noch können Staaten, Unternehmen und Bürger gegensteuern, doch sie müssen jetzt handeln und ihre Komfortzone verlassen. Einige tun es bereits.

Von **Andrea Trinkwalder**

US-Konzerne beherrschen 70 Prozent des europäischen Cloudmarktes. Microsoft dominiert zudem die Business-Software, Amazon den Handel, Google und Meta die Werbevermarktung nebst Kommunikation und Meinungsbildung – und alle zusammen die KI-Entwicklung. Die meisten europäischen Staaten haben mittlerweile erkannt, dass eine solche Abhängigkeit auf fast kompletter digitaler Ebene – Infrastruktur und Services – wirtschaftlich ziemlich ungesund ist und sie erpressbar macht. Einige Experten halten sie sogar für noch fataler als die für einige Staaten anhaltende Abhängigkeit von russischen Rohstoffen. Essenzielle IT-Dienste können exorbitant verteuert oder aus

politischen Gründen abgestellt werden – was Berichten zufolge bereits geschah. Aufgrund von Sanktionen soll Microsoft den Chefankläger des Internationalen Strafgerichtshofs Karim Khan von seinem E-Mail-Konto abgekoppelt haben; der Konzern bestreitet das. Fakt ist: Ersatz lässt sich weder kurz- noch mittelfristig organisieren, wenn US-Dienste den Stecker ziehen.

Die handliche MS-Office-Suite von gestern verzahnte sich im Laufe der Jahre mit immer mehr Organisations- und Kommunikationsprozessen im Unternehmen. Weil Firmen davon nicht mehr so einfach wegkommen, kann Microsoft die Kosten für sein Office 365 inzwischen beliebig erhöhen – und gönnte sich in diesem Jahr Preiserhöhungen von bis zu 40 Prozent. Die nach Übersee abfließenden Lizenzgebühren fehlen in der Staatskasse und schmälern den Gewinn hiesiger Unternehmen – und würgen damit auch schleichend deren Handlungs- und Innovationsfähigkeit ab. Allein an Microsoft zahlte der Bund im Jahr 2024 rund 205 Millionen Euro an Lizenz- und Servicegebühren, fast fünfmal so viel wie 2015.

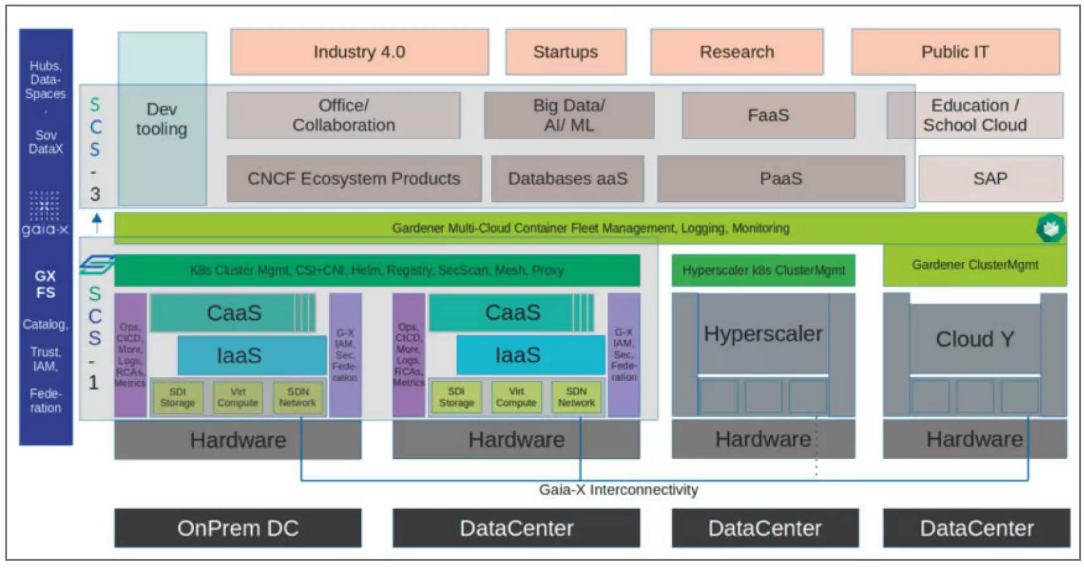
Auch angesichts des rauer werdenden Verhältnisses zu den USA dürfte spätestens 2025 klar geworden sein, dass es nicht mehr nur darum gehen darf, welche Apps und Systeme fachliche Anforderungen erfüllen. Anwender müssen auch berücksichtigen, ob man nach mehreren Jahren Nutzung noch den Anbieter wechseln kann, wenn die Lizenzgebühren zur Schutzgelderpressung mutieren. Wer sich dem allmählichen Ausbluten entziehen will, muss digital so souverän wie möglich bleiben oder werden, sprich: eine kluge, pragmatische Vermeidungs- und Exit-Strategie entwickeln. Viele haben die Weichen bereits gestellt und machen sich an die Umsetzung, wie Beispiele aus mehreren Ländern – auch Deutschland – zeigen. Der Rettungsanker heißt Open Source, von klassischer Software über Server bis hin zur künstlichen Intelligenz.

Daten unter Kontrolle

Digitale Souveränität beginnt bei den Daten. Volle Kontrolle hat nur, wer sie auf eigenen Servern hält – vorausgesetzt, er kann sie wirksam vor fremdem Zugriff schützen. Eine Alternative sind Clouddienste, die Daten europäischer Kunden auf Servern in der EU speichern und sich damit den EU-Regeln von Datenschutzgrundverordnung (DSGVO), Digital Services Act (DSA), Digital Markets Act (DMA) und KI-Verordnung (AI Act) unterwerfen.

Folgerichtig sind sensible Daten und Firmengeheimnisse derzeit bei US-Clouddiensten wie Amazon Web Services (AWS), Microsoft Azure oder Google Cloud nicht gesetzeskonform aufgehoben. Denn

Bild: SCS



Beispiel für ein IT-Ökosystem mit vier Cloudanbietern, von denen zwei den Sovereign Cloud Stack nutzen. So lassen sich auch Hyperscaler und andere inkompatible Angebote mitverwalten.

diese unterliegen dem US Cloud Act, auch wenn die Firmen eine Niederlassung in der Europäischen Union haben und die Verarbeitung in einem hiesigen Rechenzentrum zusichern. Ebendieser Cloud Act verpflichtet Microsoft & Co., Daten auf Geheiß an die US-Behörden herauszugeben. Ein richterlicher Beschluss ist nicht erforderlich und eine Informationspflicht über den Datenabfluss besteht nicht. Die Betroffenen müssen also nicht einmal erfahren, wann, in welchem Umfang und zu welchen Zwecken ihre Daten transferiert werden – was auch der staatlichen Industriespionage Tür und Tor öffnet. Dies hat der Chefjustiziar von Microsoft France bereits zähneknirschend eingeräumt.

Die EU hat sich daher schon früh entschieden, eine Basis für Plattformdienste zu schaffen und zu fördern, die das Streben nach Unabhängigkeit und Konformität zur europäischen Gesetzgebung unterstützt. Dafür hat sie 2019 das Projekt Gaia-X ins Leben gerufen.

Unter der Ägide der Open Software Business Alliance (OSBA, osb-alliance.de) entstand der Sovereign Cloud Stack (SCS, scs.community) – eine modulare, auf Open-Source-Komponenten errichtete Architektur, die es europäischen Anbietern und Selbst-Hostern erleichtert, nicht nur konform zur DSGVO zu handeln, sondern auch verschiedene Unabhängigkeits- und Interoperabilitäts-Level zu erreichen. Ein wichtiges Merkmal dieser Architektur ist, dass sie über Datenschutz und Sicherheit hinaus auch Maßstäbe für dezentrale, offene Plattformen definiert und bereits einen technischen Rahmen dafür bietet. Das ist sozusagen der Gegenentwurf zum ressourcen- und kaptialintensiven Verdrängungswettbewerb, den sich die Techkonzerne liefern.

Mit Plusserver oder SysEleven gibt es bereits einige deutsche und europäische Hoster, die ihre Services auf Basis des Sovereign Cloud Stacks anbieten; die im August gestartete Bayerische Schulcloud (ByCS) stützt sich auf SCS und Open-Source-Anwendungen wie ownCloud. Auch in der Schweiz könnte die Technik bald als Basis für die landesweite Verwaltungsplattform Swiss Government Cloud dienen. Allerdings arbeiten die Hyperscaler hart daran, den Begriff der „souveränen Cloud“ für ihre eigenen Zwecke zu kapern.

Innovative Länder

In Deutschland leistet Schleswig-Holstein mit seiner umfassenden Open-Source-Digitalstrategie Pio-

Bild: Landesregierung Schleswig-Holstein

Der digital souveräne Arbeitsplatz des Landes Schleswig-Holstein steht auf mehreren Säulen. Die Umstellung erfolgt stufenweise, beginnend mit LibreOffice und Open-XChange.

nierarbeit. Das Küstenland konzipiert seine Infrastruktur sowohl für die eigene Verwaltung als auch für Unternehmen, die mit ins Boot kommen dürfen; zum einen, weil auch die Firmen einen Weg aus der digitalen Abhängigkeit finden sollen, und zum anderen, weil die Anwendungen besser werden, je mehr Entwickler sich beteiligen.

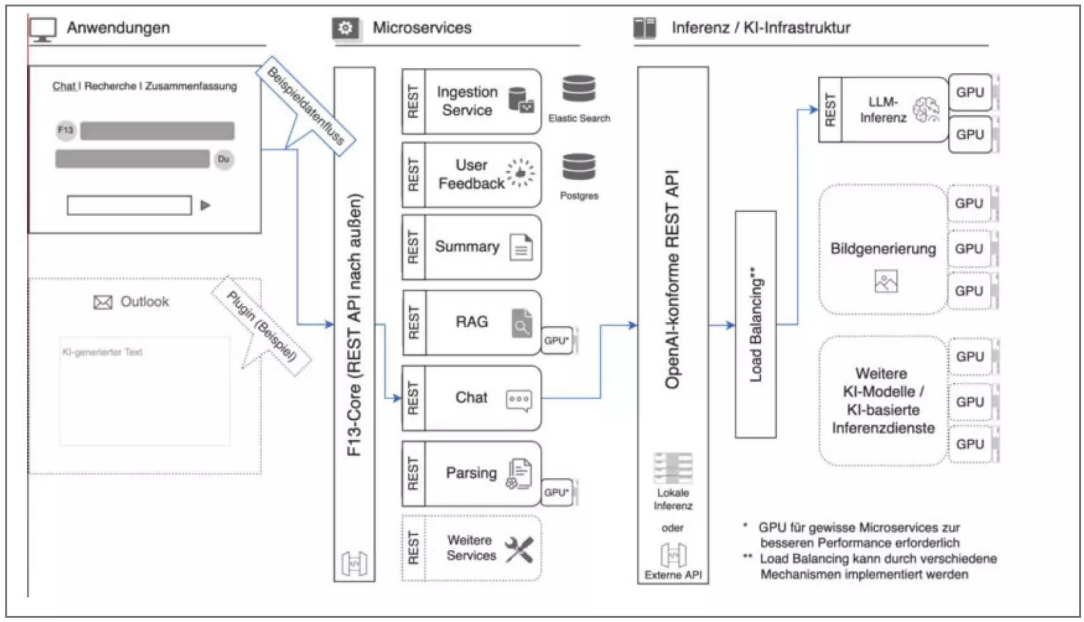
Bereits 2017 beschloss die damalige Regierung aus CDU, Grünen und FDP, künftig mehr Open Source einzusetzen. Eine umfangreiche Studie bestätigte, dass sich mit GNU/Linux ein souveräner Arbeitsplatz realisieren ließe, der die Anforderungen der Verwaltung erfüllt (siehe ct.de/wnxv). Seitdem geht es konsequent voran – aber stufenweise, um Probleme besser abfangen zu können. 2024 installierte die Landesverwaltung LibreOffice auf allen 25.000 Arbeitsplätzen und stellte auf das offene OpenDocument-Dateiformat um. Ebenfalls seit 2024 werden die Postfächer von MS Exchange auf OpenXChange migriert; 2500 pro Woche, sodass man ab Oktober 2025 auf Exchange und Outlook verzichten kann. Bis dahin sollen auch 70 Prozent der Rechner komplett ohne MS Office auskommen, damit möglichst wenige Lizenzen verlängert werden müssen.

Um auch Windows als Betriebssystem bis zum Windows-10-Ende den Rücken zu kehren, reicht die Zeit nicht, weshalb die Behörde noch Windows-11-Lizenzen kaufen musste. Deshalb fällt die souveräne Lösung während der Umstellungsphase nicht unbedingt billiger aus oder kommt sogar teurer. Langfristig gewinnt die Verwaltung aber Flexibilität und Raum für eigene Entscheidungen, und sie befreit sich aus einer Kostenspirale, die sich mit schwindenden Ausweichmöglichkeiten vermutlich immer schneller drehen wird. Das gesparte Geld soll teilweise in die Förderung von Open-Source-Softwareprojekten fließen.

Brain from „The Länd“

Sie schwätzt hoffentlich nicht nur schwäbisch, die in Baden-Württemberg entwickelte Verwaltungs-KI F13. Seit einigen Monaten ist der im Ländle auf Textarbeiten trainierte Helfer als openCode-Projekt freigegeben; über die openCode-Plattform können deutsche Verwaltungen ihre selbst entwickelten Tools als Open Source untereinander austauschen. Die zusammen mit dem Heidelberger KI-Pionier Aleph Alpha entwickelte Software soll im Wesentlichen helfen, Information zu recherchieren und Dokumente zusammenzufassen. Mittlerweile haben sich die Wege der Partner getrennt: Aleph Alpha entwickelte

Bild: f13-os.de



Der in Baden-Württemberg zusammen mit Aleph Alpha entwickelte KI-Assistent F13 soll die Arbeit in der Verwaltung erleichtern. Er kann beliebige Sprach- und andere Machine-Learning-Modelle einbinden.

eine eigene kommerzielle GovTech-Anwendung namens „Pharia Government Assistant“, während das staatliche Innovationslabor namens Innolab_bw F13 zum Open-Source-System umbaute. Seitdem ist der Verwaltungsassistent auch modellunabhängig, sodass er jedes beliebige (Open-Source-)Machine-Learning-Modell für seine Zwecke einspannen kann. Unter der Oberfläche von F13 recherchieren unter anderem wahlweise, Mixtral von Mistral oder Llama von Meta.

Ungeteilte Begeisterung scheint das Werkzeug im Behörden- und Schulalltag noch nicht zu entfachen, aber auch ChatGPT wurde nicht an einem Tag erbaut. Immerhin: Die Entwickler haben noch einige Anpassungsmaßnahmen auf der Roadmap stehen und mit dem Saarland, Hamburg und Nordrhein-Westfalen hat F13 auch direkt erste Abnehmer und Mitstreiter gefunden, die das System aktiv voranbringen. In einem Pilotprojekt werden derzeit unter anderem der Chat weiterentwickelt sowie die Recherchefunktion auf Saarland-spezifische Quellen zugeschnitten.

Zurück zu Baden-Württemberg: F13 ist kein Inselprojekt, sondern Teil der Strategie, sich als KI-Zentrum Deutschlands zu etablieren. „The Länd“ hat mit dem Cyber Valley in Karlsruhe und dem AI Center in Tübingen nicht nur bedeutende Forschungsstandorte geschaffen. Es beherbergt auch die beiden einzigen deutschen Firmen, die einigermaßen konkurrenzfähige generative Grundlagenmodelle entwickeln: Jonas Andrulis hat sich mit Aleph Alpha in Heidelberg angesiedelt und entwickelt dort das Sprachmodell Pharia (ehemals Luminous) sowie die KI-Plattform PhariaAI. Der Bildgenerator Flux wiede-

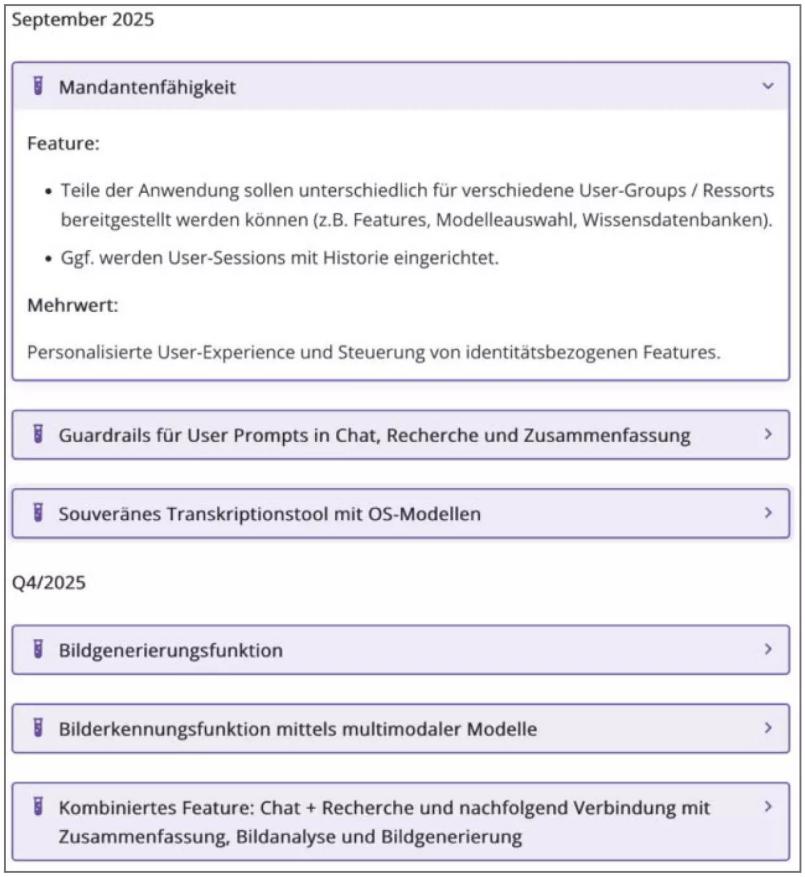


Bild: Sreenshot/f13-os.de

F13 steht seit Kurzem als Open Source zur Verfügung. Die ursprünglichen Entwickler haben selbst noch einige Verbesserungspläne; unter anderem sollen Funktionen zur Bildverarbeitung und -generierung hinzukommen.

Bild: Marco Abram/CSCS

Das Schweizer Messer von morgen: der mit Ökostrom betriebene KI-Supercomputer Alps in Lugano. Hier trainierten die Forscher der EPFL und der ETH ihr mehrsprachiges Open-Source-Sprachmodell Swiss-LLM.

rum gedeiht in Freiburg bei Black Forest Labs, das ehemalige Mitarbeiter der Stable-Diffusion-Entwicklerfirma Stability.AI gegründet haben.

Investorengelder fließen zu einem guten Teil von klassischen Konzernen aus der Region, die die Souveränität vorantreiben, schon allein aus Eigeninteresse: Schwarz Digits (die Digitalisierungstochter der Lidl-Schwarz-Gruppe) und SAP. Mit seiner Cloudinfrastruktur StackIT positioniert sich Schwarz Digits selbst als Anbieter für die souveräne Cloud; zusammen mit SAP will es zum Hyperscaler werden und damit zur Alternative für Microsoft Azure, Google Cloud und AWS. Den KI-Diensten liegt PhariaAI von Aleph Alpha zugrunde, auf dessen operatives Geschäft Schwarz Digits offenbar zunehmend Einfluss gewinnt: Um dieses kümmert sich seit Anfang des Jahres der ehemalige Schwarz-Digits-Manager Reto Spörri.

Schweizer Messer

Wer, wenn nicht die Schweiz? Das kleine Alpenland inmitten von Europa und außerhalb der EU ist so etwas wie der Inbegriff von Eigenständigkeit. Deshalb setzt es sich zwangsläufig besonders ernsthaft mit Fragen der digitalen Souveränität auseinander. Aktuell sieht sich die Schweiz ähnlich wie die EU mit Erpressungsversuchen aus Übersee in Form hoher Zölle konfrontiert.

Diese sollen dem Digitalwirtschaftsexperten Reinhard Riedl vom Institut Digital Technology Management der BFH Wirtschaft in Bern zufolge vermutlich bewirken, dass die Schweiz auf die Regulierung marktbeherrschender digitaler Plattformen verzichtet (was die Regierung bereits erwogen hatte) oder Gesetze wie das Bundesgesetz über den elektronischen Identitätsnachweis (BGEID) anpasst. Es stehen Befürchtungen im Raum, nach denen dann auch große ausländische Konzerne einmal nationale Schweizer IDs ausstellen könnten.

Den einzigen Weg dahin, wie sich die Schweiz ihre eigenen kulturellen Werte bewahren beziehungsweise deren Veränderung selbst gestalten kann, sieht Riedl in der technologischen Flucht nach vorn. Dazu wären eigene Datenräume zu schaffen und viel mehr in Open Source zu investieren, vor allem im Bereich der künstlichen Intelligenz und ganz besonders bei den großen Sprachmodellen.

Die Schweiz ist zwar klein, aber innovativ. Gemessen am Bruttoinlandsprodukt rangiert sie weltweit auf Platz 5 bei den Ausgaben für Forschung und Entwicklung, die der öffentliche und der private Sektor in etwa zu gleichen Teilen schultern. Die technischen Hochschulen EPFL in Lausanne und ETH in Zürich haben einen hervorragenden Ruf, und die Forscher dort können zudem auf ein nationales Netzwerk von 200 KI-Experten zurückgreifen. Kein Wunder, dass auch die großen Tech-Konzerne wie Meta, Google, Amazon sowie die KI-Start-ups Anthropic und OpenAI eigene Forschungs- und Entwicklungsabteilungen in Zürich aufgebaut haben.

Kein Wunder daher auch, dass die Schweiz das Selbstbewusstsein hat, ein eigenes Large Language Model zu schaffen. Es heißt Apertus, ist Open Source und konsequent vielsprachig, mit über 1.000 Sprachen trainiert. Anfang September wurde es in zwei Modellgrößen mit 8 Milliarden und 70 Milliarden Parametern veröffentlicht. Das Swiss-LLM ist das Ergebnis einer gezielten nationalen Strategie, um digitale Souveränität zu erlangen und insbesondere generative künstliche Intelligenz zu entwickeln, die sich an den eigenen Werten orientiert: unter anderem Mehrsprachigkeit, Transparenz und eine offene Diskussionskultur.

Bei Apertus soll die Öffentlichkeit daher nicht nur Einblick in den Code und die verwendeten Parameter bekommen, sondern auch in die Trainingsdaten und das entscheidende Alignment, das dem Sprachmodell Regeln für eine zivilisierte Gesprächsführung vorgibt. Mit welchen Texten wurde trainiert, mit welchen nicht und wie wurden die Daten verarbeitet?

„Die Stärken und Schwächen der Modelle haben ihren Ursprung in den Trainingsdaten, deshalb wollen wir das alles sehr transparent machen“, erklärt Martin Jaggi, Professor für Machine Learning an der EPFL und Mitglied des Swiss AI Initiative Steering Committee. Solche für Forscher und Zivilgesellschaft essenziellen Einblicke gewähren Open-Weights-Modelle wie Llama nicht und auch das chinesische Open-Source-Modell DeepSeek gibt Trainingsdaten und -modalitäten nicht preis.

Wer einen Haken sucht, findet ihn am ehesten in der allgemein schlechten Ökobilanz großer generativer Modelle, aber immerhin bemühen sich die Schweizer um einen verantwortungsvollen Umgang mit Ressourcen: Das Land betreibt in Lugano einen eigens für KI-Forschung konzipierten Supercomputer namens Alps und versorgt ihn mit 100 Prozent klimaneutralem Strom. Alps ist das Herzstück des Swiss-LLM-Projekts.

Was die Praxistauglichkeit im Vergleich zu den riesigen LLMs von OpenAI und Google angeht, macht eine Studie von Nvidia – immerhin größter Profiteur des immensen Ressourcenverbrauchs großer LLMs – Hoffnung: Die Forscher sehen in agentenbasierenden Systemen der Zukunft sogar kleinere und spezialisierte Sprachmodelle im Vorteil. Um Programmieraufgaben zu orchestrieren und auszuführen, benötige man lediglich einen Bruchteil der Fähigkeiten, die ein Universal- oder gar ein Reasoning-LLM besitzt. Über kurz oder lang komme es darauf an, dass diese auf Consumer-Hardware laufen.

Nachholbedarf beim Bund

In Deutschland mangelt es leider an einer einheitlichen Strategie des Bundes, die sich in verbindlichen Regeln oder Gesetzen manifestiert. So kommt es, dass jeder US-Hyperscaler in Deutschland mittlerweile eine vermeintlich „souveräne Cloud“ anbietet, obwohl aufgrund des Cloud Act klar ist, dass sie nicht einmal die Mindestanforderungen der DSGVO erfüllen können, nämlich persönliche Daten oder Firmengeheimnisse dem Zugriff der US-Behörden zu entziehen. Unreflektierte Aktionen von Bundesbehörden und -politikern verschärfen das Problem, beispielsweise die jüngst vereinbarte Kooperation des Bundesamts für Sicherheit in der Informationstechnik (BSI) mit Google zum Aufbau einer sicheren, „souveränen“ Cloudinfrastruktur für öffentliche Verwaltungen.

Die Gesellschaft für Informatik kritisiert diese scharf: „Die geplante Zusammenarbeit des BSI mit Google gefährdet nicht nur unsere nationale Sicherheit und Cybersecurity, sondern auch die strategische Autonomie und Handlungsfähigkeit der Bundesbehörden und erhöht die Erpressbarkeit von Bundesregierung, Bundesverwaltung und der deutschen Bürgerinnen und Bürger“, schreibt sie in einer Stellungnahme. „Eine Partnerschaft mit einem Unternehmen, dessen Kerngeschäft die Monetarisierung von Daten ist, wirft erhebliche Bedenken hinsichtlich der Unabhängigkeit und Souveränität der geplanten Cloudlösung auf.“

Auch diesbezüglich kann man sich einiges von der Schweiz abschauen, die zumindest einige der Gaia-X-Prinzipien bundesweit verbindlich vorschreiben will, und zwar mit dem Bundesgesetz über den Einsatz elektronischer Mittel zur Erfüllung von Behördenaufgaben, kurz EMBAG. Es verpflichtet die Bundesbehörden zu Open Source Software und Open Government Data by Default, wenn sie Software beschaffen oder selbst entwickeln. Damit ist der Sovereign Cloud Stack nach Meinung einiger Experten geradezu prädestiniert als Fundament für die „Swiss Government Cloud“, für die demnächst die Ausschreibungen formuliert werden. Verankert man SCS-Anforderungen in der Ausschreibung, so die Hoffnung der Parlamentarischen Gruppe Digitale Nachhaltigkeit, blieben die Hyperscaler von sich aus fern.

Was Privatleute tun können

Auch Privatleute können ihren Teil gegen die Dominanz der US-Konzerne beitragen, etwa indem sie gemeinnützige, genossenschaftlich organisierte sowie regionale Projekte unterstützen – sei es mit Spenden oder als Nutzer. Über den lokalen Buchhandel bekommt man Bücher mittlerweile schneller geliefert als von Amazon und auch für andere Waren gibt es genügend Alternativen zum unübersichtlich gewordenen globalen Marktplatz, die oft auch besser sortiert sind.

Ungleich schwieriger gestaltet es sich, Kommunikationsdienste wie WhatsApp und andere Netzwerke zu verlassen, die den Vendor-Lock-in perfektioniert haben. Eigentlich sollen DSA und DMA die Anbieter – insbesondere Meta – verpflichten, ihre Plattformen zu öffnen, sodass man sich beispielsweise auch via Signal oder Threema mit WhatsApp-Freunden austauschen kann. Doch ähnlich wie die Hyperscaler die „souveräne Cloud“ interpretiert auch Meta die „Interoperabilität“ auf eine sehr eigentümliche beziehungsweise eigennützige Weise, weshalb die XMPP Standards Foundation (XSF) Alarm schlägt.

Die XSF entwickelt seit 25 Jahren den offenen Standard eXtensible Messaging and Presence Protocol, den übrigens auch Google und Meta intern als Basis ihrer proprietären Messenger nutzen. Der von Meta an die EU-Kommission übermittelte Vorschlag zu Messenger-Interoperabilität wecke ernsthafte Bedenken, dass der Konzern mithilfe dieser Maßnahmen sogar noch mehr Kontrolle über Nutzerdaten erlangen könne, warnt die XSF. Insbesondere dass Meta den Mitbewerbern mithilfe von Verschwiegenheitsklauseln und anderen Vereinbarungen nur einen restriktiven Zugang gewähren möchte und nach außen auf Schnittstellen statt offene Protokolle (wie XMPP) setzt, werde seine Position als Gatekeeper stärken.

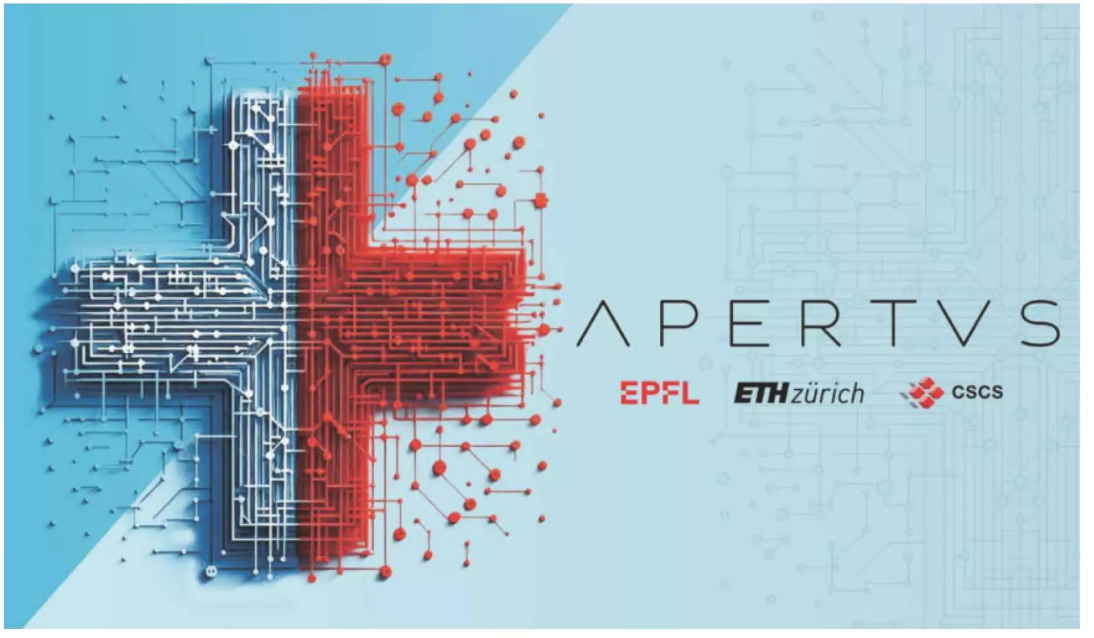
EU-Bürger, -Unternehmen und -Institutionen werden nur dann eine echte und freie Wahl haben, wenn sich sowohl die Kommission als auch die Wettbewerbshüter nicht von solchen technischen Taschenspielertricks blenden und über den Tisch ziehen lassen.

Studien und Quellen:
ct.de/wnxv

Machen statt hoffen

Deutschland verkauft Autos, China produziert sie, die USA versorgen uns mit IT, billige Energie kommt aus Russland: Diese praktische globale Arbeitsteilung, die EU-Länder reich machte, hat sie in eine fatale Abhängigkeit manövriert. Die wohlstandssichernde Arbeitsteilung löst sich auf, und während es zu deutschen Autos massenhaft Alternativen gibt, mutieren die essenziellen IT-Dienste zur tickenden Zeitbombe mit hohem wirtschaftlichen und politischen Erpressungspotenzial.

Nun lässt sich weder kurz- noch mittelfristig alles ersetzen, was in Übersee jahrzehntelang an Code gereift ist. Aber es gibt eine Basis, auf der man aufbauen kann, um das Ruder wieder in Richtung eines gesunden Wettbewerbs herumzureißen: Open Source sowie eine eigene KI- und IT-Infrastruktur, die europäische Werte stärkt, statt sie zu unterminieren. Unsere Beispiele aus der Schweiz und Deutschland zeigen, dass das machbar ist. (atr) ct



Alternativen aus der Schweiz

Vielsprachigkeit, Transparenz, Respekt vor geistigem Eigentum und ein Hauch Lokalkolorit: Das konsequent offene große Sprachmodell aus den führenden Schweizer KI-Schmieden verinnerlicht europäische Werte.

Von **Andrea Trinkwalder**

David gegen Goliath, Gallier gegen Römer, Frodo gegen Mordors Mächte. Seit jeher faszinieren Erzählungen, wie die vermeintlich Schwächeren die Übermächtigen zu Fall bringen. Nun, in der modernen Fassung: das offene Schweizer Sprachmodell Apertus gegen ChatGPT, Gemini, Llama & Co. Das eine finanziert mit Schweizer Forschungsgeldern und Gesetzestreue by Design, trainiert auf einem mit Ökostrom betriebenen Supercomputer in den Alpen. Die anderen gebaut auf Investorenmilliarden und ohne Rücksicht auf Mensch und Umwelt.

Stolz präsentierten Schweizer Forscher der ETH Zürich, des EPFL aus Lausanne und des CSCS in Lugano Anfang September ihren Gegenentwurf zur künstlichen Intelligenz US-amerikanischen Zuschnitts: ein großes, multilinguales Sprachmodell mit 70 Milliarden Parametern, das die Urheberrechte respektiert sowie Offenheit und Transparenz zur Maxime erhoben hat. Obgleich für mittelgroße LLMs

wie Apertus nicht die strengsten Anforderungen der europäischen KI-Verordnung gelten, haben sich die Wissenschaftler rund um Projektleiter Martin Jaggi diesen freiwillig unterworfen. Damit liefern sie den ersten Proof of Concept für ein DSGVO- und AI-Act-konformes Sprachmodell.

Es ist unter den großen Sprachmodellen eines der wenigen reinen Open-Source-Projekte, die alles Wesentliche komplett offenlegen: nicht nur die Trainingsgewichte wie Metas Llama (Open Weights) oder den Quellcode nebst Gewichten wie das chinesische Deepseek. Apertus dokumentiert auch die verwendeten Trainingsdaten, die Art, wie diese gefiltert wurden, die Checkpoints der Trainingsläufe, die Finetuning- und Alignment-Prozeduren und mehr. Kurzum, der aus dem Lateinischen stammende Name ist Programm und wer nachvollziehen möchte, wie das Sprachmodell funktioniert und warum es bestimmte Antworten gibt oder wirres Zeug erzählt, der kann alle möglichen und unmöglichen Ursachen bis hinab zu den Trainingsdaten untersuchen.

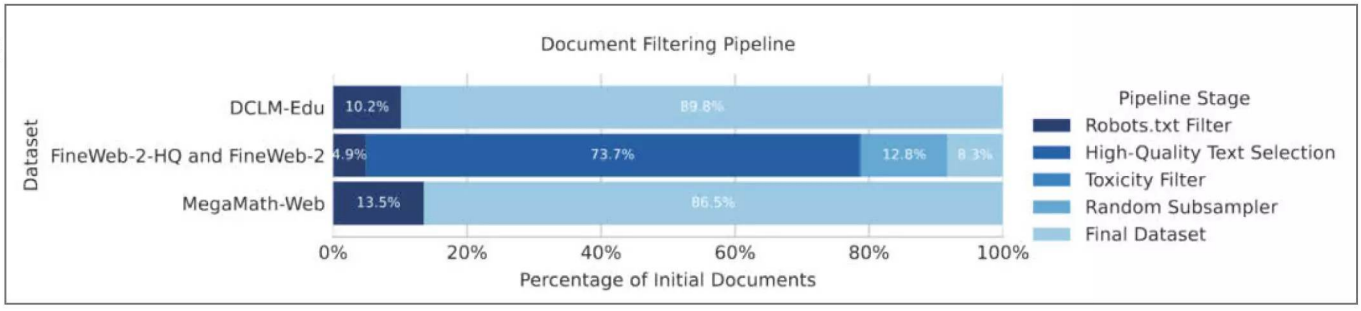
Dank permissiver Apache-Lizenz lässt sich die KI auch mit eigenen Daten von Grund auf neu trainieren (Pretraining) oder nachjustieren (Post-Training). Wie sich die Sprachfähigkeit des Basismodells verbessern lässt, haben die Forscher selbst anhand von Schweizerdeutsch und Rätoromanisch demonstriert, wie ein Sprecher der ETH Zürich gegenüber c't erläutert: „Die Tatsache, dass wir auf über 1000 Sprachen trainiert haben, bedeutet nicht automatisch Konversationsfähigkeit in allen diesen Sprachen. Jedoch erfordert es nun signifikant weniger Aufwand, die Performance in diesen Sprachen zu verbessern. Mit wenigen tausend Post-Training-Samples verbesserte sich die Performance erheblich."

Transparent, offen, nachvollziehbar

Insbesondere für die Wissenschaft, aber auch für besonders sensible Anwendungen, sind solche souveränen, transparenten Modelle von unschätzbarem Wert, weil die kommerziellen Anbieter mittlerweile keine Details mehr über ihre Architekturen veröffentlichen – und schon gar nicht ihre Trainingsdaten offenlegen. Letzteres hat wettbewerbliche und urheberrechtliche Gründe. Denn OpenAI, Google, Meta & Co. haben nicht nur Unmengen an Material aus dem Internet zusammengetragen, sondern auch einen eigenen wertvollen Fundus an Nutzerdaten angehäuft, mit denen sie ihre Modelle optimieren. Viele Urheber, darunter Zeitungsverlage und Künstler, möchten ihre Inhalte aber gar nicht als KI-Trainingsmaterial zur Verfügung stellen, einige gehen juristisch dagegen vor. Den Nachweis zu führen, gelingt nur auf Umwegen; das Material nachträglich aus den Modellen zu entfernen, ist bei den proprietären Anbietern quasi unmöglich.

Die ETH- und EPFL-Forscher stützen das Pretraining, also das Basistraining für den allgemeinen Spracherwerb, auf mehrere, teils multilinguale Standard-Trainingsdatensammlungen wie FineWeb. Letzteres besteht aus Inhalten von Websites, Blogs et cetera, die von Web-Crawlern zusammengetragen und in bestimmten Zeitabständen aktualisiert werden. Es handelt sich also um Schnappschüsse des

Bild: M. Jaggi et al.



Beim Schweizer Sprachmodell Apertus ist alles offengelegt, vom Quellcode über die Architektur bis hinab zur Filter-Pipeline für die Trainingsdaten.

LLMs im Vergleich

Large Language Model	Apertus	Olmo 2	DeepSeek (V3/R1)	Llama 4
Hersteller, URL	Swiss AI, swiss-ai.org	Ai2, allenai.org	DeepSeek, deepseek.com	Meta, llama.com
Modellgrößen (Mrd. Param.)	8B, 70B	1B, 7B, 13B, 32B	671B, kleinere Destillate (Mixture of Experts)	109B, 400B (Mixture of Experts)
Open Source	✓	✓	–	–
Open Weights	✓	✓	✓	✓
Trainingsdaten offengelegt	✓	✓	–	–
Trainings-Pipeline offengelegt	✓	✓	–	–

✓ ja – nein

WWW, die zu unterschiedlichen Zeitpunkten angefertigt wurden.

Urheber können in der robots.txt ihrer Website hinterlegen, ob die KI-Crawler ihre Inhalte verarbeiten dürfen oder nicht. Einige halten sich daran, andere nicht. Unabhängig davon besteht das Problem, dass etwa Künstler oder Autoren, die ihren Willen erst jetzt ausdrücklich via robots.txt bekunden, die bis dato gescrapten Inhalte nicht mehr aus den Sammlungen herausbekommen. Die Schweizer Forscher haben deshalb einen Filtermechanismus ausgetüftelt, mit dem sie nachträgliches Opt-out realisieren. Sie fragen die aktuellen Einwilligungserklärungen ab und bereinigen anhand dieser Informationen die kompletten Korpora der vergangenen Jahre, im Falle von FineWeb von 2013 bis 2024.

Der Verzicht auf urheberrechtlich möglicherweise bedenkliches Material hat einen Preis, wie die Forscher ebenfalls dokumentieren. Dadurch gehen etwa 8 Prozent der englischen und 4 Prozent der multilingualen Daten verloren. In Standard-Benchmarks wie HellaSwag, die auf Allgemeinwissen prüfen, erreichte Apertus dennoch ungefähr den Stand der Llama-3.1-Modelle; Einbußen sind einer separat durchgeführten Studie zufolge vor allem beim Erkennen falscher Thesen zu erwarten.

Sicherheitsrisiko Auswendiglernen

Dass LLMs bisweilen dazu neigen, Trainingsdaten wortwörtlich wiederzukäuen, ist ein Problem. Nicht nur, weil sie sich damit den zweifelhaften Ruf eines „Statistical Parrot“ – also eines nachplappernden, aber nicht verstehenden Papageien – eingehandelt haben, sondern weil es mitunter rechtliche Konsequenzen hat, wenn eine KI Passagen aus Büchern oder Zeitungsartikeln abschreibt (Copyright-Verstöße), die Krankheitsgeschichte real existierender Person veröffentlicht (Datenschutz) oder Firmengeheimnisse ausplaudert (Security).

Um solche Pannen zu verhindern, wenden LLM-Entwickler üblicherweise spezifische Alignment-Verfahren an. Das Alignment ist ein wichtiger und sehr aufwendiger Teil der zweiten Trainingsstufe (Post-Training). Mithilfe großer, aufwendig kuratierter Datensätze lernen die Textgeneratoren dabei unter anderem, sich angemessen auszudrücken, keine Anleitungen für kriminelle Handlungen zu liefern, etwa den Bomben- oder Waffenbau, oder auch Aufgaben zu strukturieren (Reasoning). Weil sich solche Sicherheitsmechanismen aber mit geschicktem Prompting oder gezielten Finetuning-Attacken umgehen lassen, wie Forscher anhand von Gemini 1.5, Llama 2 und GPT-4 zeigten, erprobte das Apertus-Team ein Trainingsverfahren namens Goldfish Loss, das bereits im Pretraining ansetzt, also in der ersten Stufe. Mithilfe einer gezielten Maskierungsstrategie bringt es die Fehlerfunktion dazu, das wörtliche Reproduzieren zu vermeiden. Dadurch soll das Modell ein robusteres Verhalten entwickeln und grundsätzlich vom Original abweichend formulieren.

Zauber des Anfangs

Jedem Anfang wohnt ein Zauber inne. Bei großen Sprachmodellen währt der genauso lange, bis man sie testet. Auch Apertus startet mit den üblichen Schwächen wie Halluzinationen, veraltetem Wissen et cetera, die man von anderen LLMs kennt. Die sind, nicht überraschend, ausgeprägter als bei den etablierten, die ihre Erstlinge bereits seit Jahren mit Daten aus dem täglichen Gebrauch verfeinern können.

Auch Apertus braucht noch einigen Feinschliff. Einen Teil davon könnte es während der diesjährigen Swiss AI Weeks bekommen haben. Im Rahmen der

fünfwöchigen Veranstaltungsreihe im September erhielten Teilnehmer des Hackathons Zugriff auf das API und ein Chat-Interface. Teamleiter Martin Jaggi betont außerdem, dass die jetzt veröffentlichte Version nicht als Ende, sondern als Startpunkt zu sehen ist. Das LLM wird von EPFL/ETH weiterentwickelt, konkret geplant seien spezifische Modelle für Medizin, Recht und den Finanzsektor. Zu Funktionserweiterungen in Richtung Multimodalität, Reasoning-Funktionen, Coding und RAG gebe es Überlegungen, aber noch keine konkreten Pläne.

Schon jetzt leisten die Schweizer aber einen wertvollen Beitrag zur viel beschworenen menschenzentrierten, werteorientierten KI. Anstatt schulterzuckend geltendes (EU-)Recht zu ignorieren und Rechtsbrüche mit Verweis auf einen höheren Wert ihrer Technik zu rechtfertigen, suchen Jaggi und seine Mitstreiter nach Gestaltungsmöglichkeiten, berücksichtigen die teils gegenläufigen gesellschaftlichen Interessen, dokumentieren Verfahren sowie deren Stärken und Schwächen ausführlich und legen damit den Grundstein für eine konstruktive Entwicklung und Erforschung der künstlichen Intelligenz. Bei der Industrie stoße Apertus auf sehr großes Interesse, das Feedback sei überwiegend positiv. Vor allem Entwickler von Produkten und Services, die LLMs in agentischen Systemen oder KI-Workflows einsetzen, sähen in der konsequenten Transparenz und Compliance einen erheblichen Wert, so die erste Bilanz der ETH Zürich.

Wer Apertus ausprobieren möchte, kann dies am bequemsten über das Public AI Inference Utility, ein von Mozilla gegründetes und von der Metagov-Community finanziertes Open-Source-Projekt. Modelle für die lokale Installation bekommt man via Hugging Face, derzeit in Trainingspräzision mit 70 Milliarden und 8 Milliarden Parametern sowie in quantisierten Versionen für Apples MLX-Framework. Mit der Ollama-Umgebung gibt es noch ein Kompatibilitätsproblem zu lösen, GGUF-Versionen dafür seien in Arbeit. Eine Demo mit Retrieval Augmented Generation, also der Recherche in eigenen Dokumenten, findet sich bei ZüriCityGPT. Der Dienst wertet die Informationen auf der Website der Stadt Zürich aus. (atr) c't

Downloads und Quellen:
ct.de/w8k3

(Bild: Ulrike Weis/KI/heise medien)

Zwischen Ermittlung und Überwachung

Die Polizeien in Hessen, Bayern und Nordrhein-Westfalen ermitteln mit Software des US-Unternehmens Palantir, Baden-Württemberg zieht nach. Rechtlich steht der Einsatz auf wackligem Fundament.

Von **Andrea Trinkwalder**

Spionage, Militär, Polizeiarbeit – das ist das Metier der US-Firma Palantir und ihrer Analysesoftware Gotham. Auch einige deutsche Bundesländer haben Gotham angeschafft, um ihre Ermittler im Kampf gegen schwere Kriminalität zu unterstützen. Es gebe nichts Vergleichbares, lautet das einhellige Argument, weshalb man auf diese Software angewiesen sei. Doch nicht nur das Produkt ist umstritten, auch der Gründer des Unternehmens Alex Karp sowie dessen größter Investor Peter Thiel genießen einen zweifelhaften Ruf. Thiel hält Freiheit und Demokratie für unvereinbar und Karp hält Überwachung für ein probates Mittel, um bürgerliche Freiheiten zu bewahren. „Software“, schreibt Karp, „ist so sehr wie alles andere ein Produkt der rechtlichen und moralischen Ordnung, der sie entstammt, und spielt eine Rolle in ihrer Verteidigung.“

Zudem stehen die Umstände, unter denen die bundesdeutschen Gotham-Ableger namens HessenData, VeRA (Verfahrensübergreifende Recherche

und Analyse) in Bayern und DAR (Datenbankübergreifende Recherche und Analyse) in Nordrhein-Westfalen zum Einsatz kommen, rechtlich auf wackligem Fundament.

Bürgerrechtler haben nun Klage erhoben, bevor bundesweit Fakten geschaffen werden. Denn auch das Bundesland Baden-Württemberg und Bundesinnenminister Alexander Dobrindt (CSU) für die Bundesebene liebäugeln mit dem Ermittlungswerkzeug, das Daten zusammenführt und Analysen beschleunigen soll. Die Thematik ist komplex und wirft auf mehreren Ebenen Probleme auf. Nicht zuletzt drängt sich angesichts der aktuellen Beziehungen zu den USA eine Grundsatzfrage in den Vordergrund: Wäre es nicht klug, bei der inneren Sicherheit digitale Souveränität anzustreben, solange es noch geht?

Wir beleuchten das gesamte Spektrum: von sicherheitspolitischen Erfordernissen und rechtlichen Grenzen über den Charakter und die Arbeitsweise der Palantir-Analytics-Systeme bis hin zu der Frage, ob es (staatlichen) Auftraggebern nicht eigentlich egal sein kann, was CEOs und Investoren von Software-Unternehmen politisch so denken.

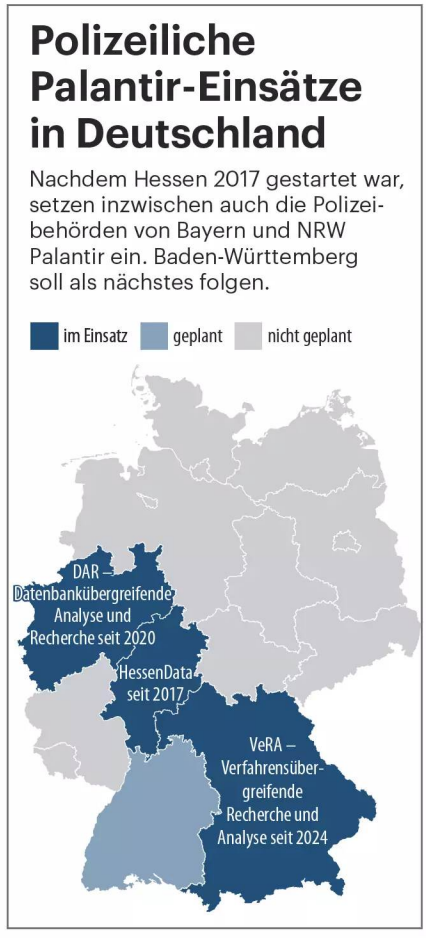
Wie Palantir tickt

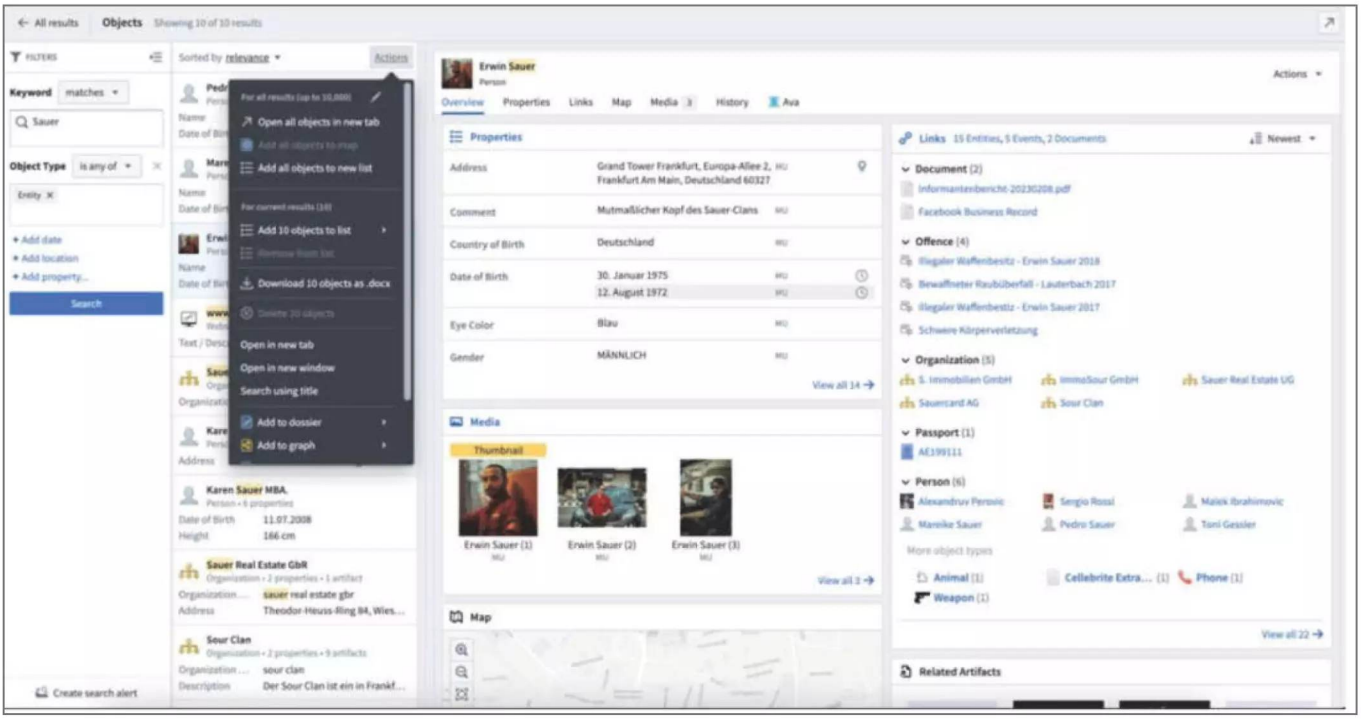
Palantir wurde 2003 gegründet, unter anderem von Alex Karp und PayPal-Gründer Peter Thiel. Als technisches Fundament diente PayPal-Software: Die Idee war, die für den Zahlungsdienstleister entwickelten Datenanalyse- und Betrugserkennungsalgorithmen so weiterzuentwickeln, dass sie Staaten dabei helfen, Terrorismus zu bekämpfen und militärische Ziele aufzuspüren. Heraus kam Gotham: eine Plattform, mit der sich sehr große, verteilte und unstrukturierte Datenbestände rasch verknüpfen und analysieren ließen, speziell zugeschnitten auf die Ermittlungs- und Aufklärungsarbeit staatlicher Akteure.

Wie raffiniert und einzigartig die zugrunde liegenden Algorithmen tatsächlich sind und in welchem Ausmaß künstliche Intelligenz (KI) beziehungsweise Machine Learning enthalten sind, darüber gibt es viele Spekulationen und wenig Belastbares. Vermutlich ist der Unterbau von Gotham vergleichbar mit handelsüblicher Analytics- oder Business-Intelligence-Software. Fakt ist aber, dass es Palantir geschafft hat, eine Oberfläche zu bauen, mit der Praktiker in den Büros sowie im Einsatz rasch und intuitiv arbeiten können. Kriminalbeamte oder Soldaten können es einfach benutzen, ohne sich fortgeschrittene IT-, Statistik- oder Programmierkenntnisse aneignen zu müssen.

Hellseherische Ordnung

Dreh- und Angelpunkt dieses Statistikpakets für Aufklärung und Überwachung ist die sogenannte Ontologie, also das semantische Datenmodell. Dieses soll sämtliche notwendigen Arbeitsabläufe digital abbilden, von der einfachen Personenabfrage über die erweiterte Recherche im direkten Umfeld bis hin zur strategischen Einsatzplanung und -steuerung. Die Ontologie muss für jede Branche und jedes Unternehmen eigens angepasst werden, was während der Einführungsphase der Software in Zusammenarbeit mit Palantir-Mitarbeitern geschieht; für die hiesigen Landeskriminalämter ist die deutsche Niederlassung Palantir Technologies GmbH mit Sitz in Frankfurt am Main zuständig.





(Bild: Palantir)

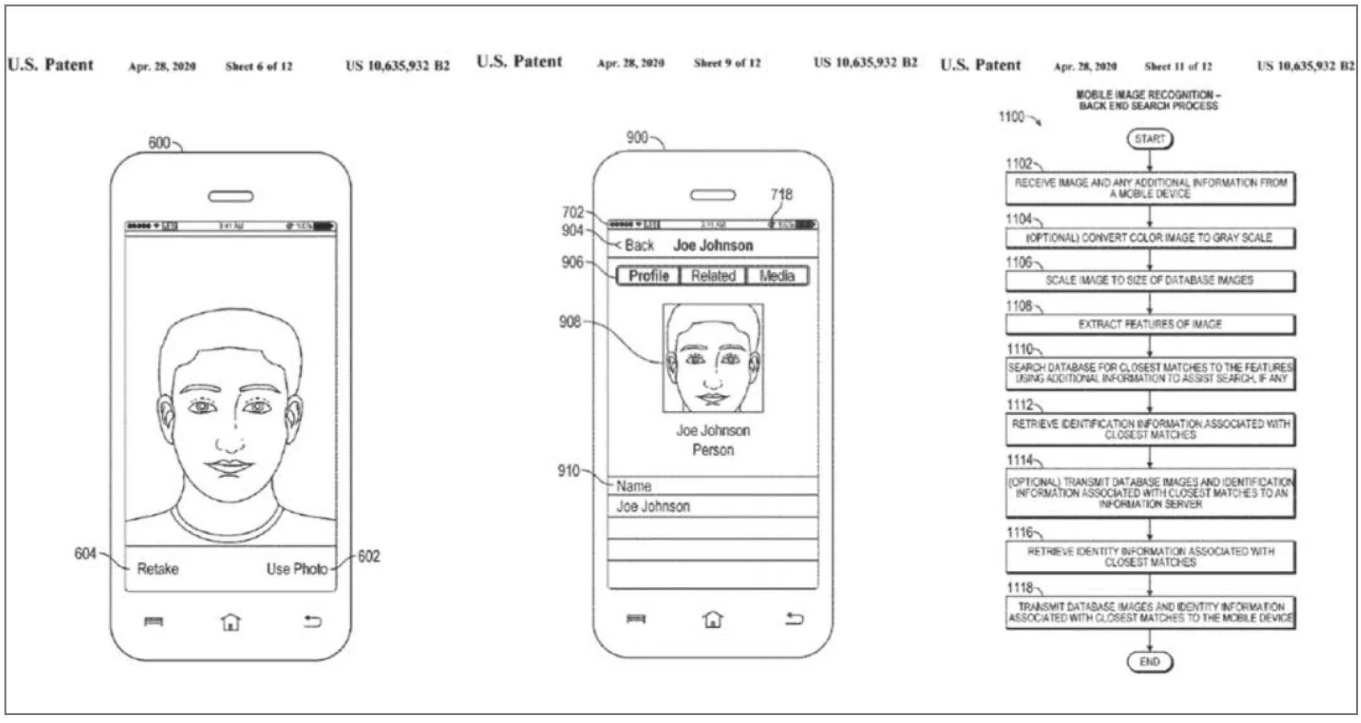
Die Ermittlungsplattformen der Landeskriminalämter von Hessen, Bayern und Nordrhein-Westfalen beruhen auf Palantir Gotham. Die Analyseplattform verknüpft und visualisiert Informationen aus unterschiedlichen Datenbanken und macht sie leicht durchsuchbar.

Für diese Anpassung müssen die Beteiligten sämtliche für die Ermittlungen relevanten Datentypen mitsamt ihren Eigenschaften beziehungsweise Metadaten definieren – und im Idealfall alle Möglichkeiten, wie die Objekte miteinander in Beziehung stehen können: etwa Personen untereinander oder Personen mit (Tat-/Wohn-)Orten, Handys, Fahrzeugen, Waffen, älteren Fällen et cetera. Im Vorfeld gilt es also, sämtliche Verknüpfungen anzulegen, die sich potenziell als relevant etwa bei der Fahndung nach Terroristen oder der Bekämpfung von Bandenkriminalität erweisen könnten.

Zusätzlich zu den konkreten Merkmalen der Objekte selbst werden auch alle möglichen Prozess- und Interaktionsparameter hinterlegt sowie Zugriffsrechte vergeben: Wie erscheint eine bestimmte Person im User Interface, welche Aktionen kann man mit ihr ausführen (überwachen, auf Karte anzeigen, privates Netzwerk überprüfen et cetera), welche Nutzer dürfen auf das Objekt zugreifen beziehungsweise es editieren. Diese Methode der Datenverknüpfung firmiert in der IT unter dem Begriff „Knowledge Graph“. Anders als klassische relationale Datenbanken können solche Graphen auch dynamische, sich verändernde oder kausale Beziehungen abbilden, im besten Fall in Echtzeit. Diese Kunst der Datenorganisation beherrscht Palantir vermutlich nicht besser als andere; auch Amazon und Google wenden sie in ihren Rechenzentren an.

Aber Palantir hat seine Daten-Pipeline – anders als das Gros der eher Social-Media-affinen Tech-Branche – von Anfang an auf die Bedürfnisse von Militär und Polizei hin optimiert. Genau diese maßgeschneiderte Echtzeitanalyse umfangreicher, verteilter Datenmengen ist es, womit sich Palantir in Sicherheitskreisen ins Spiel bringt und zunehmend auch bei anderen großen Institutionen und Firmen. Im Jahr 2015 veröffentlichte es mit Foundry einen auf Unternehmen zugeschnittenen Gotham-Ableger. Damit bekam Palantir einen Fuß in die Tür von

(Bild: Palantir)



Über Recherchen im Patent-Portfolio versuchten Forscher herauszufinden, wie Palantir-Software Wissen verarbeitet und Vorhersagen trifft. Ein Patent beschreibt die Social-Media-Recherche anhand von Handyfotos.

großen Firmen und zivilen Organisationen wie der Food and Drug Administration (Tracking von Covid-19-Daten und Impfstoff-Lieferketten) oder des britischen National Health Services (NHS); seit 2020 ist Palantir an der Börse.

Agenda inside

Damit die auf unterschiedliche Datenquellen verteilten Informationen korrekt miteinander verknüpft werden, müssen sie zunächst in die Ontologie gemappt werden. Nicht alle liegen sauber in Tabellen und Datenbanken vor; wichtige Details finden sich zudem in unstrukturierter Form in Protokollen, Fotos, Videos und Metadaten. Daten aufzubereiten und zu harmonisieren, ist also oft noch mit viel Handarbeit verbunden. Es lässt sich zwar einiges automatisieren, zunehmend auch mithilfe von großen Sprachmodellen, dennoch bleibt der Aufwand groß. Auch dies dürfte ein Grund dafür sein, warum die Einführung von Gotham-Systemen so lange dauert und eine enge Zusammenarbeit mit Palantir-Mitarbeitern erfordert.

In welchem Ausmaß Machine-Learning-Algorithmen dabei helfen, Daten aufzubereiten und in das Schema der Ontologie einzupassen, oder anfangs dazu gedient haben, die Basis-Ontologie zu entwickeln, ist nicht bekannt. Dennoch entsteht insgesamt ein komplexer Algorithmus, in dem es auch blinde Flecken oder Verzerrungen geben kann, in der Art und Weise, wie er arbeitet und Zusammenhänge herstellt. Und, wie die Forscher Andrew Iliadis und Amelia Acker in einer tiefgehenden Analyse der Patente von Palantir zu bedenken geben: Palantir arbeitet Daten für die Ontologie nicht neutral auf, um die Welt zu beschreiben, wie sie ist, sondern wie das Unternehmen sie sieht: „Palantir will, dass seine Überwachungsplattform weltweit das De-facto-Betriebssystem für Regierungen und Unternehmen wird, die mit den sensibelsten Daten arbeiten.“ [1].

Rechtlicher Rahmen gefordert

Die Kritik, die sich am Einsatz der Gotham-Varianten entzündete, findet auf mehreren Ebenen statt. Zum einen geht es darum, ob man dem Unternehmen und seiner Software angesichts der demokratiefeindlichen Agenda, die ihre Gründer verfolgen, überhaupt trauen kann. Eng damit verknüpft ist das Streben nach digitaler Souveränität und insbesondere der Unabhängigkeit von den USA; zu beidem später mehr. Ganz grundsätzlich gilt es aber zunächst den rechtlichen Rahmen zu gestalten, innerhalb dessen eine solche Ermittlungssoftware im Bund und den Ländern eingesetzt werden darf.

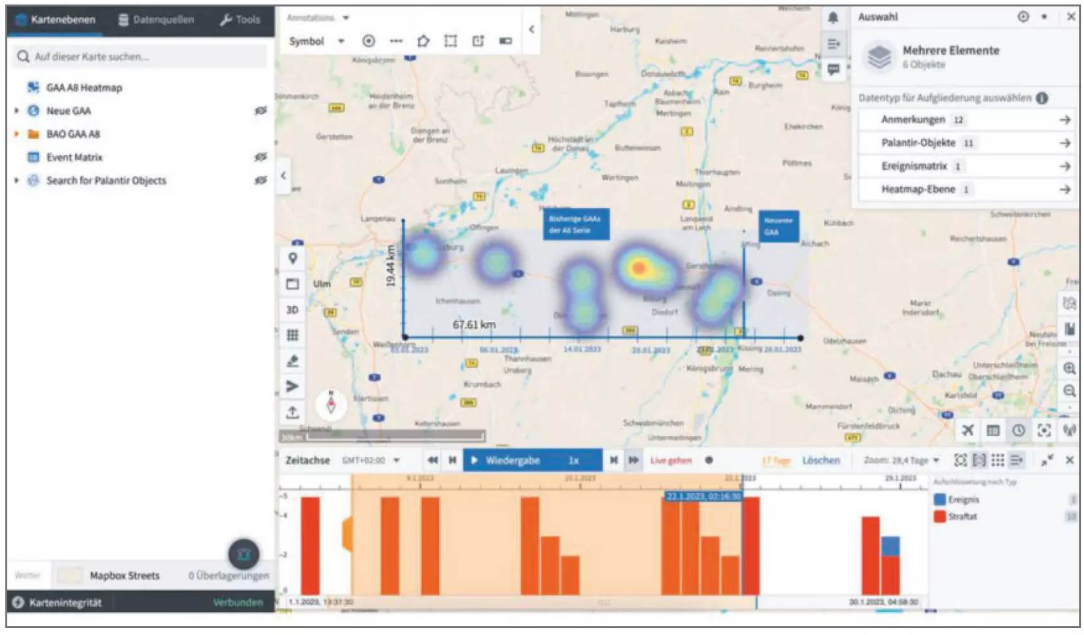
Welche Datenbestände miteinander verknüpft werden dürfen und welche nicht, wie umfassend Ermittler das Umfeld eines Verdächtigen ausleuchten dürfen und bei welchen Vergehen die Polizei damit arbeiten darf: Solche entscheidenden Details hatte zum Beispiel das Bundesland Hessen in seinem extra überarbeiteten Landespolizeigesetz nicht ausreichend geregelt, als es sein Palantir-gestütztes Ermittlungssystem HessenData in Betrieb nahm, urteilte das Bundesverfassungsgericht (BVerfG) 2023 [2]. Die obersten Verfassungshüter skizzierten auch eine Art Korridor, in dem sich digitalisierte und zu einem gewissen Grad automatisierte Ermittlungsarbeit bewegen könnte.

Grundsätzlich stellten sie klar: Persönlichkeitsrechte, Datenschutz & Co. verdonnern die Polizei keineswegs dazu, digital hochgerüstete Verbrecher mit Klemmbrett oder anderen veralteten Systemen zu jagen. Ermittler dürfen getrennte Datenbanken zusammenführen und diese automatisiert auswerten. Ebenso wenig gibt es ein generelles Verbot von Machine-Learning-Verfahren. Aber: Der Staat darf nicht bei jedem geringfügigen Delikt eine komplette, breit angelegte Durchleuchtungs- und Überwachungsmaschinerie anwerfen, nur weil er die Technik dafür besitzt.

Denn nach Ansicht des BVerfG sind algorithmische Analysemethoden perspektivisch dazu geeignet, durch Zusammenführen und Auswerten unterschiedlicher Datenquellen eine verfassungswidrige Rundumüberwachung zu bewirken. Eine solch intensive Überwachung verletze die Menschenwürde, wenn sie „sich über einen längeren Zeitraum erstreckt und derart umfassend ist, dass nahezu lückenlos alle Bewegungen und Lebensäußerungen des Betroffenen registriert werden und zur Grundlage für ein Persönlichkeitsprofil werden können“. Es gilt hier also, effektive Begrenzungen zu finden,

(Bild: Palantir)

Die intuitiv bedienbare Oberfläche führt Ermittler schneller zum Ergebnis. Recherchen, die früher mehrere Tage gedauert haben, verkürzen sich auf wenige Minuten, berichten LKA-Mitarbeiter.



(Bild: Palantir)

Zusätzliche Module wie etwa eine Kartendarstellung helfen bei der Einsatzplanung und beim Beobachten verdächtiger Personen.

um diese Form der Datenauswertung in Einklang mit dem Grundgesetz zu bringen. Konkret sieht das BVerfG den Gesetzgeber in der Pflicht, eine Systematik zu entwickeln, die die Verhältnismäßigkeit der Eingriffe gewährleistet, etwa indem er Eingriffsschwellen definiert oder die Datenverarbeitungsmethode einschränkt.

Diese Anforderungen erfüllten Hessen und Bayern mit ihren zu schwammig und offen formulierten Polizeiaufgabengesetzen nicht, als sie Palantir in Betrieb beziehungsweise Testbetrieb nahmen. Sie handelten also mindestens teilweise verfassungswidrig und besserten ihre Rechtsgrundlage offenbar nur unzureichend nach. Weder das nachträglich aktualisierte hessische noch das bayerische Polizeiaufgabengesetz genügten den Anforderungen, die das BVerfG formuliert, monierte die Gesellschaft für Freiheitsrechte (GFF) und legte erneut Verfassungsbeschwerde ein: gegen die mangelnde gesetzliche Grundlage von HessenData bereits im vergangenen Jahr, gegen das BayPAG im Juli dieses Jahres. Auch der bayerische Landesdatenschutzbeauftragte Thomas Petri sowie der Rechtswissenschaftler Mark Zöller von der Ludwig-Maximilians-Universität München hatten sich bereits kritisch geäußert und die unzureichenden Schutz- und Kontrollmechanismen gerügt (PDFs via ct.de/wcwk). Ihr Hauptkritikpunkt ist, dass im Rahmen von Ermittlungen unter Umständen ein zu großer Personenkreis regelmäßig automatisiert überprüft würde. Denn die Archive, die in VeRA und HessenData derzeit eingebunden sind, enthalten auch eine große Menge Daten unbeteiligter Personen, etwa aus Funkzellenabfragen oder aus den Vorgangsverwaltungssystemen der Behörden. In letzteren sind beispielsweise Angaben zu Zeugen, Hinweisgebern oder Personen gespeichert, die Anzeige erstattet haben. Allein dadurch steigt die Wahrscheinlichkeit für unbescholtene Bürger, doch einmal einem verdächtigen Muster zu entsprechen. Noch ungleich höher ist die Wahrscheinlichkeit für Personen, die aufgrund eines Bias im Datenmodell oder in einem Verarbeitungsalgorithmus des Systems tendenziell eher als verdächtig eingeschätzt werden, etwa aufgrund ihrer Herkunft oder ihres Wohnorts.

Bereits in seinem 2022 veröffentlichten 32. Tätigkeitsbericht hatte der Bayerische Landesdatenschutzbeauftragte seine rote Linie definiert: „Die Anwendung von VeRA ist grundsätzlich auf solche Daten zu beschränken, die die Polizeibehörden unter besonderen Voraussetzungen für Zwecke der vorbeugenden Gefahrenabwehr speichern. Eine Einbeziehung

(Bild: Palantir)

Zu den ersten Palantir-Kunden gehörte der militärische Sektor. Mit den Fähigkeiten, die sogenannte Kill Chain (Angriffskette) zu modellieren – also Ziele und Zielpersonen aufzuspüren und diese zu eliminieren – wirbt das Unternehmen offensiv.

von Daten aus der polizeilichen Vorgangsverwaltung IGVP, die einen enormen Umfang hat und größtenteils Daten von unbescholtenen Bürgern enthält, lehne ich ab."

Außerdem entziehe sich die Software weitgehend der öffentlichen Kontrolle. Weder seien regelmäßige, verpflichtende Sicherheitsüberprüfungen noch Inspektionen durch unabhängige Stellen wie etwa den Landesbeauftragten für Datenschutz vorgesehen.

Freiwillige Selbstkasteiung

Die Landeskriminalämter in Nordrhein-Westfalen, Hessen und in Bayern betreiben ihre Gotham-Ableger derzeit in einem reduzierten Modus, der weit von dem entfernt ist, was die Überwachungs- und Einsatzplanungsplattform kann. Derzeit seien ausschließlich Datenbanken der Polizei angebunden, KI-Funktionen würden nicht genutzt und externe Quellen ebenfalls nicht durchforstet; überhaupt haben die Plattformen keinen Zugriff auf das Internet. Demzufolge wäre ein Profiling mithilfe öffentlicher Social-Media-Accounts etwa nach Gesichtern, Vorlieben oder Netzwerken derzeit nicht möglich – zumindest nicht mit VeRA, DAR und HessenData.

Aber aus den Behördenangaben zur Nutzung von VeRA & Co. geht auch hervor, dass die Plattformen nicht nur ausschließlich dann eingesetzt wurden, wenn eine „konkrete Gefahr für hochrangige Rechtsgüter" vorlag, wie es verfassungsrechtlich konform wäre, sondern auch bei geringfügigeren Anlässen wie etwa „drohender Gefahr" oder Raub. In Hessen hat zudem ein großer Personenkreis Zugriff, nämlich über 2000 Kriminalbeamte, die mit ihr im letzten Jahr über 14.000 Ermittlungen durchgeführt haben.

Diese großzügige Ausweitung des Nutzerkreises und der Anwendungsbereiche zeigen, dass es klarer gesetzlicher Regeln, vorgeschriebener Kontrollen und eines konsequenten Monitorings solcher Systeme bedarf.

Aktuelle politische Lage

Bleibt die Frage, ob Gotham das Mittel der Wahl sein sollte. Denn angesichts der aktuellen politischen Lage in den USA rückt ein lange verdrängter Aspekt wieder in den Vordergrund: die digitale Souveränität. Geplant war und ist noch immer, eine eigene, länderübergreifende Ermittlungs- und Datenplattform zu entwickeln. Doch das Projekt namens P20 geriet immer wieder ins Stocken und wird erst in einigen Jahren einsatzbereit sein. Der Wunsch nach einem Arbeitsmittel, das seinen Namen verdient, ist also nachvollziehbar.

Dennoch ist auffällig, wie eilig die jeweiligen Bundesländer Gotham angeschafft haben, und das hat möglicherweise auch mit der Verkaufsstrategie des Herstellers zu tun. Die Innenminister der Länder begründeten die Anschaffung der Palantir-Software meist mit der Notwendigkeit, die Gefahr von Terroranschlägen frühzeitig zu erkennen und diese zu verhindern. Wie effektiv die Plattformen tatsächlich sind, müsste aber unabhängig untersucht werden, was bisher nicht geschah. So bleibt es bei anekdotischer Evidenz und einer sehr dünnen Datenlage. Ob und in welchem Umfang diese Systeme an den proklamierten

Ermittlungserfolgen beteiligt waren, lässt sich von der Öffentlichkeit genauso wenig prüfen wie Datensicherheit, Fehlerquoten oder systematische Verzerrungen. Details über die Funktionsweise der Systeme bleiben wahlweise aus ermittlungstaktischen Gründen unter Verschluss, oder um das Geschäftsgeheimnis des Softwareherstellers zu wahren.

Aus Sicht des bayerischen Datenschutzbeauftragten lieferte die Sicherheitslage schon 2022 kein überzeugendes Argument für VeRA: „Die Sicherheitslage scheint sich nahezu jährlich zu verbessern. So wurde für die Polizeiliche Kriminalstatistik (PKS) im Jahr 2021 in Bayern die niedrigste Kriminalitätsbelastung seit 44 Jahren und gleichzeitig die höchste Aufklärungsquote seit 27 Jahren vermeldet. [...] dies führt schlussendlich zu der Fragestellung, ob angesichts dieser Sachlage noch ein derart eingriffsintensives Instrument wie VeRA zusätzlich erforderlich ist und wenn ja, wie der Einsatz verhältnismäßig ausgestaltet werden kann."

Aktuell drängt Bundesinnenminister Alexander Dobrindt darauf, die Plattform auch bundesweit einzuführen (Bundes-VeRA). Denn Bayern hatte für VeRA bereits einen Rahmenvertrag ausgehandelt, dem sich auch andere Bundesländer anschließen können. Das baden-württembergische, CDU-geführte Innenministerium unterzeichnete im März dieses Jahres hastig einen 25 Millionen Euro teuren Fünfjahresvertrag für Gotham – ebenfalls ohne klare, vorab definierte Regeln und ohne Zustimmung des Koalitionspartners, den Grünen. Hauptgrund für die Eile war eine auslaufende Preisbindung; danach hätte das Land das Doppelte zahlen müssen. Die zunächst protestierenden Grünen knickten rasch ein, weil sie der CDU im Gegenzug Zugeständnisse beim Nationalpark Schwarzwald abringen konnten.

Bürgerrechtsorganisationen wie die GFF befürchten, dass aus der Interims- eine Dauerlösung wird und sich die Polizeien von Bund und Ländern in eine Abhängigkeit begeben, aus der sie nicht mehr so leicht herauskommen. In Bayern hat diese Einführungsphase zwei Jahre gedauert, die Ontologie wurde nach fachlichen Vorgaben des bayerischen LKA von den Palantir-Systemintegratoren umgesetzt und während der Projektierung verfeinert, wie ein BLKA-Sprecher gegenüber c't erklärte.

Berichten zufolge kann es sich allerdings schwierig und zeitaufwendig gestalten, die innerhalb der Software definierten Strukturen und aus den Daten abgeleiteten Erkenntnisse auf andere Systeme zu migrieren. Das BLKA sagte dazu gegenüber c't: „Die polizeilichen Daten der Plattform lassen sich in den gängigen Standardformaten nach Einzelfallprüfung problemlos exportieren. Ein Wechsel des Systems würde zu keinem Datenverlust führen." In welchen Standardformaten sich etwa die Ontologie exportieren ließe, erfuhren wir nicht.

Demokratiefeindlich by Design?

Bleiben noch die Elefanten im Raum, Palantir-CEO Alex Karp und Hauptinvestor Peter Thiel, und die damit verknüpfte grundsätzliche Frage, ob man Gotham überhaupt als neutrales Stück Code betrachten kann, das unabhängig von den Machtphantasien ihrer Urheber existiert und wirkt.

Gotham spiegelt die Haltung ihrer Gründer wider. Es ist auf flächendeckende und tiefgreifende Überwachung ausgelegt. Je mehr Datenquellen angebunden werden, umso nützlicher wird es. Aus Erfahrungen und Studien im Einsatz bei US-amerikanischen Polizeibehörden geht hervor, wie der anfangs beschränkte Kreis der Datenbanken immer stärker erweitert wurde – ganz einfach, weil es ging, weil sie vorhanden waren und weil sie potenziell noch mehr Erkenntnisgewinn versprachen. Außerdem stellten Forscher fest, dass sich die Arbeitsweise der Ermittler veränderte, weg vom schwerpunktmäßigen Recherchieren, hin zum Monitoring und Reagieren auf Alarmmeldungen, die vom System ausgelöst werden [3].

Derweil greift Palantir selbst regelmäßig jede sich bietende Gelegenheit auf, um klarzustellen, dass es keine aktive Rolle einnimmt, wenn Gotham zur Überwachung und Verfolgung von Immigranten durch die US-amerikanische Einwanderungs- und Zollbehörde (ICE) oder zum Markieren militärischer Ziele eingesetzt werde. Palantir sei kein KI-Unternehmen, weder an Firmen- beziehungsweise Behördendaten interessiert noch habe es Einfluss auf Projekte oder Operationen seiner Kunden.

Das mag sein. Aber während der engmaschigen Projektierungsphasen gewinnt der Dienstleister wertvolle Informationen über die Arbeits- und Denkweise seiner Kunden. Und über Semantik und Architektur seiner Plattform versucht er, diese zu beeinflussen – ganz im Sinne der Agenda von CEO Alex Karp.

Der bayerische Innenminister Joachim Herrmann (CSU) zumindest sah darin keinen gravierenden Konflikt. Begründung: Wenn der bayerische Staat den Neubau eines Gebäudes beauftrage, spiele die Gesinnung des Bauunternehmers schließlich auch keine Rolle. Man habe durchaus Bauchschmerzen, aber leider gebe es nun mal nichts Besseres. (atr) c't

Literatur

[1] Andrew Iliades, Amelia Acker, The seer and the seen: Surveying Palantir's surveillance platform, The Information Society, 38(5), 334-363: heise.de/s/drLw0

[2] BGH, Urteil vom 16. 2. 2023, Az. 1 BvR 1547/19 , 1 BvR 2634/20: heise.de/s/oPMWV

[3] Sarah Brayne, Big Data Surveillance: The Case of Policing, American Sociological Review, 82(5), 977-1008: heise.de/s/BPEpe

Quellen und Studien:
ct.de/wcwk

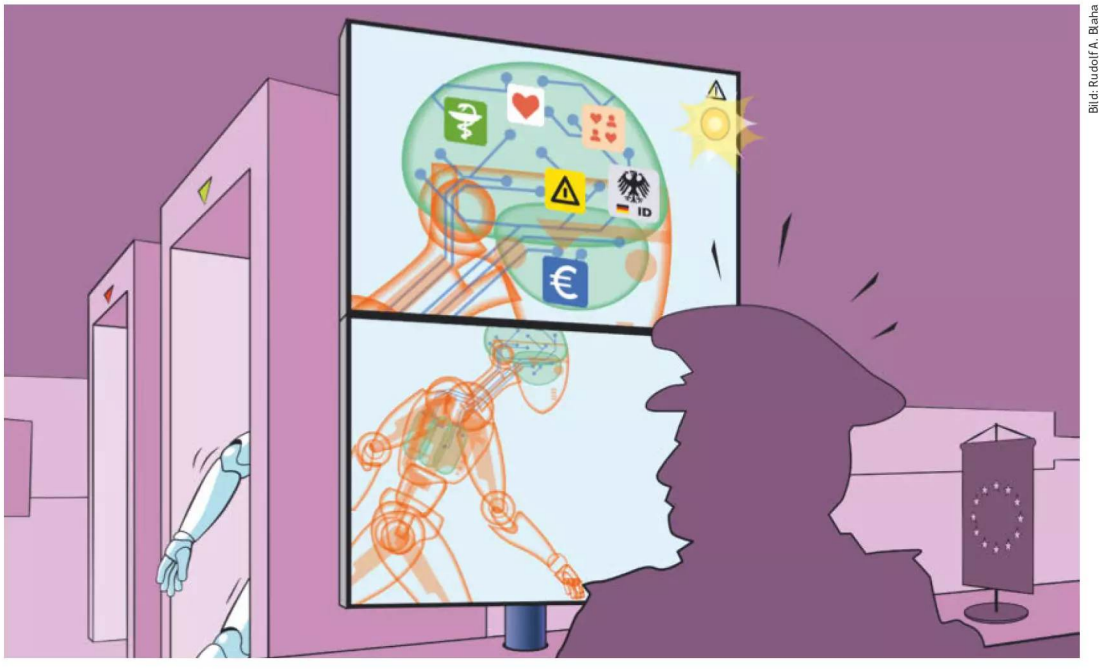


Bild: Rudolf A. Blaha

EU stellt Regeln für generative KI scharf

Seit August 2025 gilt die KI-Verordnung auch für allgemeine KI-Modelle. Im Juli erschien der „Code of Practice" mit verbindlichen Anleitungen, wie die Pflichten umzusetzen sind. Auf die Anbieter kommt viel Arbeit zu, Nutzer könnten bald von der neuen Transparenz profitieren.

Von **Holger Bleich**

Diesseits und jenseits der europäischen Grenzen beobachteten Experten mitten im Sommer hektische Betriebsamkeit bei den EU-Institutionen, nationalen Regierungen und großen Anbietern von KI-Systemen. Kurz danach, am 2. August 2025, wurde das wichtigste Kapitel der im August 2024 in Kraft getretenen EU-KI-Verordnung wirksam, in dem es um die Regulierung von KI geht, die eine „erhebliche allgemeine Verwendbarkeit" aufweist, also die in der Verordnung sogenannte „General Purpose AI" (GPAI).

Gemeint sind Modelle und Systeme, die aus dem Alltag vieler Millionen geschäftlicher und privater Nutzer kaum noch wegzudenken sind, wie Sprach-KIs, Bildgeneratoren und Musikkomponiermaschinen. Deren Anbieter sitzen bis auf wenige Ausnah-

men in den USA, beispielsweise OpenAI (GPT), Anthropic (Claude), Google (Gemini) oder Meta (Llama). Anders als kleinere europäische Anbieter wie das französische Mistral trifft sie die KI-Verordnung mit voller Breitseite.

Zur Erinnerung: Die KI-Verordnung verfolgt einen risikobasierten Ansatz. Sie stuft KI-Anwendungen gemäß ihrem Gefährdungspotenzial in Kategorien ein, von „minimal“ über „begrenzt“ und „hoch“ bis „inakzeptabel“. Je höher das Risiko ist, desto strenger sind die Auflagen. Zunächst hatte die EU-Kommission vor allem spezifische Anwendungsfälle im Fokus, etwa KI in Medizin, Verkehr oder Überwachung.

Allgemein einsetzbare Modelle wie das Ende 2022 plötzlich erfolgreiche ChatGPT passten nicht recht in dieses Schema, da sie keinem festen Einsatzgebiet zuzuordnen sind und als Baustein vieler verschiedener Systeme dienen können. Erst gegen Ende des Gesetzgebungsprozesses führte die EU daher eine eigene Kategorie für solche Allzweckmodelle (GPAI) ein. Diese Modelle gelten grundsätzlich als begrenztes Risiko, unterliegen also vor allem Transparenzpflichten.

Im angeflanschten GPAI-Regelteil (Kapitel V, Art. 51 bis 56 KI-Verordnung) schuf der Gesetzgeber sogar eine zusätzliche Unterkategorie für die mächtigsten Vertreter dieser KI-Modelle: „GPAI-Modelle mit systemischem Risiko“. Damit sind „die fortschrittlichsten KI-Modelle“ gemeint, die aufgrund ihrer Leistungsfähigkeit großflächige gesellschaftliche Schäden verursachen könnten, etwa wenn sie gefährliches Wissen leicht zugänglich machen oder sich schwer kontrollieren lassen. Die KI-Verordnung geht von einer fürs Modelltraining nötigen Rechenleistung von mindestens 10^{25} FLOPs aus. Für solche großen Modelle gelten verschärfte Vorschriften, von denen später im Artikel noch die Rede sein wird.

Kompetenzpflichten

EU-Verordnungen gelten in jedem Mitgliedsland unmittelbar als Gesetz. Aufgrund dieser Wucht sieht der Gesetzgeber stets Umsetzungsfristen vor, weshalb verschiedene Teile zu unterschiedlichen Zeitpunkten wirksam werden, je nach Erfüllungsaufwand. Die erste dieser Schwellen überschritt die KI-Verordnung am 2. Februar dieses Jahres. Die Verbote von unzulässigen KI-Systemen, etwa denjenigen, die Social Scores errechnen sollen, gelten seitdem.

Außerdem trifft Unternehmen seit dem 2. Februar die Kompetenz- und Schulungspflicht aus Art. 4 der KI-Verordnung. Anbieter und Betreiber von KI-Systemen müssen demnach sicherstellen, dass ihr Personal und andere Personen, die in ihrem Auftrag mit dem Betrieb und der Nutzung von KI-Systemen befasst sind, über eine ausreichende KI-Kompetenz verfügen. Insbesondere hat die KI-Verordnung Grundrechte, Gesundheit und Sicherheit im Blick. Mitarbeiter sollen KI-Systeme sachkundig nutzen und deren Potenziale und Risiken erkennen können.

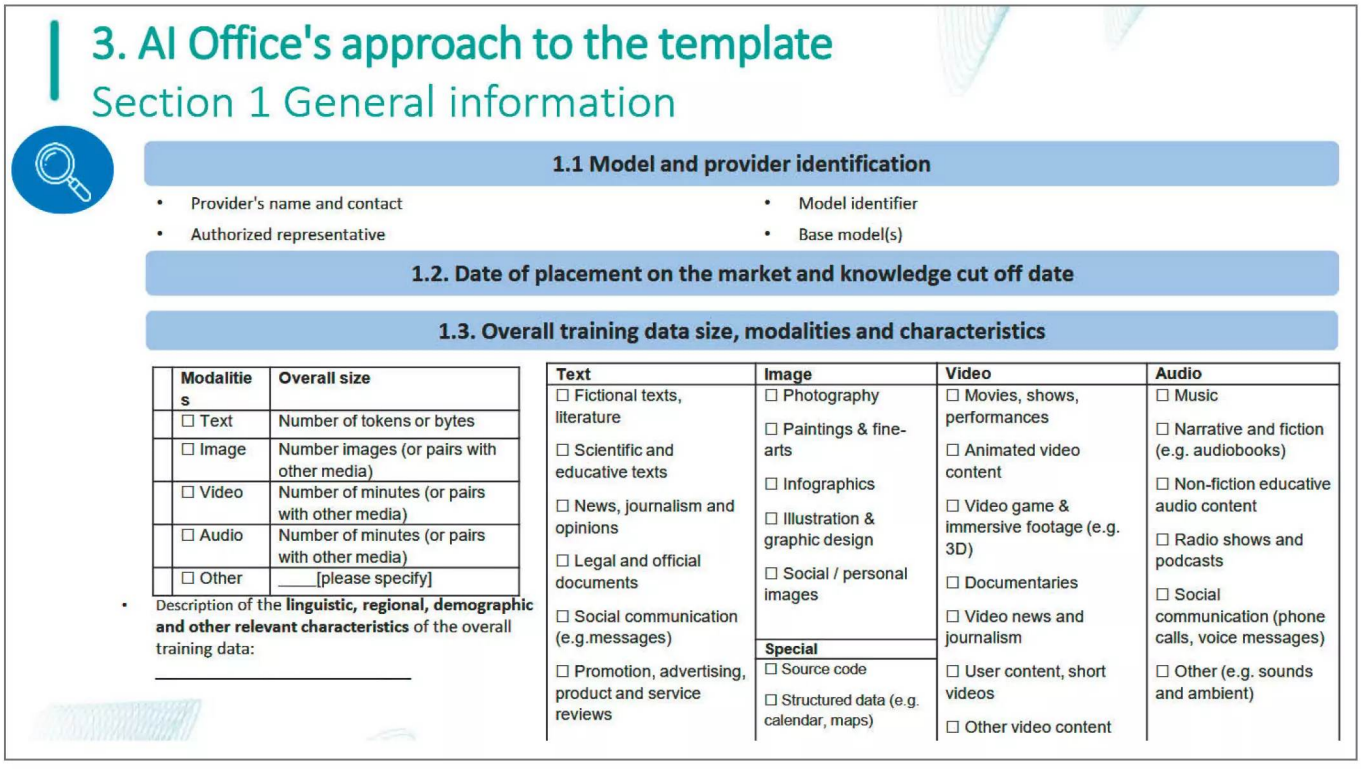
Genau deshalb schießen seit Jahresbeginn Anbieter von KI-Onlineschulungen wie Pilze aus dem Boden, die ihre Schulungen Unternehmen feilbieten. Wieder einmal schafft eine EU-Verordnung völlig neue Geschäftsmodelle. Dabei ist ein Verstoß gegen die Kompetenzpflicht in der KI-Verordnung nicht einmal mit Bußgeld sanktioniert. Juristen warnen allerdings: Wenn es durch Fehler eines Mitarbeiters zu Schäden aufgrund der Nutzung von KI kommt, könnte sich die Haftung verschärfen, falls keine vorherigen Schulungsmaßnahmen im Sinne des Artikels 5 der KI-Verordnung nachgewiesen werden können.

Regeln für Anbieter

Der wohl wichtigste Stichtag der KI-Verordnung war der 2. August 2025, als das gesamte GPAI-Kapitel voll wirksam wurde. Dies trifft Anbieter und Betreiber von KI-Modellen und -Systemen. Schon hier wird es kompliziert, denn wer ist eigentlich Anbieter und wer Betreiber?

Die Verordnung sieht es so: Anbieter ist eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System oder ein KI-Modell mit allgemeinem Verwendungszweck entwickelt oder entwickeln lässt und es unter ihrem eigenen Namen oder ihrer Handelsmarke in Verkehr bringt. Betreiber dagegen sind natürliche oder juristische Personen, die ein KI-System in eigener Verantwortung verwenden. Generell ausgenommen ist die ausschließlich private Nutzung, die nicht unter die KI-Verordnung fällt. Pflichten treffen vor allem die Anbieter, weniger die Betreiber von KI-Systemen. Im Folgenden fassen wir für Sie die ab August geltenden Vorschriften zusammen.

Ein Anbieter muss eine ausführliche technische Dokumentation seines Produkts erstellen und aktuell halten. Diese Dokumentation soll unter anderem Informationen zur Funktionsweise, den Limitierungen, Leistungsmerkmalen und Testmethoden des Modells enthalten. Die EU hoffe, dass diese Transparenz den Nutzern hilft, Risiken besser abzuschätzen und das Modell verantwortungsvoll einzuset-



3. AI Office's approach to the template

Section 1 General information

1.1 Model and provider identification

- Provider's name and contact
- Authorized representative
- Model identifier
- Base model(s)

1.2. Date of placement on the market and knowledge cut off date

1.3. Overall training data size, modalities and characteristics

	Modalities	Overall size
	☐ Text	Number of tokens or bytes
	☐ Image	Number images (or pairs with other media)
	☐ Video	Number of minutes (or pairs with other media)
	☐ Audio	Number of minutes (or pairs with other media)
	☐ Other	_____[please specify]

- Description of the **linguistic, regional, demographic and other relevant characteristics** of the overall training data:

_____

Text	Image	Video	Audio
☐ Fictional texts, literature	☐ Photography	☐ Movies, shows, performances	☐ Music
☐ Scientific and educative texts	☐ Paintings & fine-arts	☐ Animated video content	☐ Narrative and fiction (e.g. audiobooks)
☐ News, journalism and opinions	☐ Infographics	☐ Video game & immersive footage (e.g. 3D)	☐ Non-fiction educative audio content
☐ Legal and official documents	☐ Illustration & graphic design	☐ Documentaries	☐ Radio shows and podcasts
☐ Social communication (e.g.messages)	☐ Social / personal images	☐ Video news and journalism	☐ Social communication (phone calls, voice messages)
☐ Promotion, advertising, product and service reviews	**Special** ☐ Source code	☐ User content, short videos	☐ Other (e.g. sounds and ambient)
	☐ Structured data (e.g. calendar, maps)	☐ Other video content	

Bild: EU-Kommission

In einer Präsentation zeigte das AI-Büro ein Template, mit dem die Anbieter den Transparenzpflichten zu Trainingsdaten nachkommen sollen.

zen. Die Dokumentation muss der Anbieter nachgelagerten Resellern oder Betreibern weitergeben. Außerdem muss er sie auf Nachfrage den Aussichtsbehörden vorlegen.

Damit verknüpft ist eine Transparenzpflicht bezüglich der Trainingsdaten. Der Anbieter eines GPAI-Modells muss veröffentlichen, welche Inhalte er für das Training verwendet hat, zumindest in Form einer „ausreichend detaillierten Zusammenfassung". Das heißt nicht, dass er jede Quelle auflisten muss. Die Nutzer und Aufseher sollen einen Eindruck bekommen, woraus das Modell gelernt hat. Nach Ansicht der EU schafft das Vertrauen und ermöglicht Dritten, einzuschätzen, ob das Trainingsmaterial eventuell verzerrende oder problematische Daten enthielt.

Nicht zuletzt enthält das GPAI-Kapitel Vorschriften zu Urheberrechten. Die EU verlangt, dass Anbieter eine Policy einführen, um EU-Urheberrecht und verwandte Schutzrechte einzuhalten. Der Trainingsprozess und die Nutzung des Modells dürfen nicht einfach geschützte Werke verletzen. Beispielsweise muss erläutert sein, wie ein Anbieter während des Trainings seines Modells mit urheberrechtlich geschütztem Material umgegangen ist. Zudem soll er verhindern, dass generierte Inhalte unerlaubte Kopien geschützter Werke enthalten.

Systemische Risiken

Zusätzliche Pflichten gelten ebenfalls seit dem 2. August für GPAI-Modelle mit „systemischem Risiko". Der Anbieter eines solchen Modells oder Systems muss regelmäßig selbst untersuchen, welche Gefahren von der Nutzung ausgehen. Beispielsweise könnte er prüfen, ob das Modell gefährliches Wissen (wie Anleitungen zum Waffenbau) preisgeben

könnte. Entdeckte Risiken soll er mit technischen Maßnahmen reduzieren, etwa mithilfe zusätzlicher Filter oder notfalls durch menschliche Überwachung.

Außerdem gelten Anforderungen an die Cybersicherheit, also die Pflicht zu Maßnahmen gegen innere und äußere Angriffe sowie Meldepflichten für Zwischenfälle: Schwerwiegende Vorfälle oder Missbrauchsmöglichkeiten, die erst nach dem Marktstart bekannt werden, muss der Anbieter dokumentieren und den Aufsichtsbehörden melden. Sollte also ein Modell im realen Betrieb unerwartet großen Schaden anrichten, indem es beispielsweise massenhaft Falschinformationen generiert, die zu realen Schäden führen, ist der Anbieter gefordert, dies zu erfassen, weiterzugeben und abzustellen.

Eine Ausnahme von vielen dieser Pflichten gilt für die Open-Source-Community: Veröffentlicht ein Anbieter sein KI-Modell unter einer freien und offenen Lizenz inklusive aller Parameter und notwendigen Informationen, dann entfallen einige Dokumentationspflichten. So muss er dazu keine technische Doku an Behörden und nachgelagerte Anbieter übermitteln. Nicht ausgenommen sind allerdings die großen Modelle mit systemischem Risiko: Hier bleibt der Anbieter in der Pflicht, Risiken zu evaluieren und zu mindern, bevor er das Modell frei zugänglich macht. Diese Sonderregeln dürften nur bei echten Open-Source-Projekten greifen, nicht aber den teiloffenen Open-Weight-Modellen wie Metas Llama.

Code of Practice

Große Kritik erntete die KI-Verordnung dafür, dass sie von unscharfen Rechtsbegriffen strotzt und noch dazu wenig konkrete Hinweise gibt, wie das Regelwerk von Anbietern umgesetzt werden soll. Auch die nicht rechtsverbindlichen Erwägungsgründe zur Verordnung, in denen der Gesetzgeber erläutert, wie denn diese oder jene Pflicht zu verstehen sei, bietet an vielen Punkten nur wenig Hilfe.

Bei der Kritik fällt oft unter den Tisch, dass die EU-Kommission und später im Gesetzgebungspro-

zess auch das Parlament und der Rat auf das 2008 etablierte „New Legislative Framework“ (NLF) setzten. Dieses insbesondere für Produktregulierung ersonnene Prinzip hat sich die EU in einer Richtlinie selbst aufgegeben. Es geht davon aus, dass sich Märkte und Produkte ständig weiterentwickeln, weshalb Gesetze flexibel sein sollten.

Die konkrete technische Ausgestaltung soll sich nicht im Gesetz selbst, sondern in Rechtsakten, Verhaltenskodizes und Leitlinien finden. Diese erarbeitet die EU-Kommission anhand der vom Gesetz vorgegebenen Leitplanken zusammen mit den sogenannten „Stakeholdern“, also allen, die dazu etwas zu sagen haben könnten oder betroffen sind. Man nennt dies kurz Ko-Regulierung.

Wesentlich beim Erarbeiten dieser konkreten Compliance-Standards ist die Organisationsstruktur, die die KI-Verordnung definiert: Art. 64 etabliert das sogenannte AI-Büro als in Brüssel angesiedelte, supranationale Kompetenzbehörde, die in der EU-Kommission (hier Generaldirektion CNECT) angesiedelt ist. Art. 65 hebt einen AI-Vorstand aus der Taufe, der mit Vertretern der Aufsichten aus den Mitgliedsstaaten besetzt ist und das Einhalten der KI-Verordnung auf höchster Ebene überwachen soll. Außerdem existieren noch ein Beratungsgremium und ein wissenschaftlicher Beirat.

Die entscheidende Rolle kommt derzeit dem AI-Büro zu. Art. 56 KI-Verordnung definiert einen Zeitrahmen, in dem dieses AI-Büro einen möglichst konkreten „Code of Practice“ erarbeiten sollte, an den sich Anbieter und Betreiber von GPAI-Systemen halten können. Darin finden sich anders als in der Verordnung die konkreten Details, also rechtsverbindliche Leitlinien, wie sich die KI-Verordnung umsetzen lässt. Zur Ausarbeitung dieses Verhaltenskodexes darf das AI-Büro gemäß Art. 56 Abs. 3 auch die Anbieter selbst einbeziehen.

Mehr als 1000 Stakeholder waren seit der Auftaktveranstaltung am 11. Dezember 2024 am Entstehen des GPAI-Verhaltenskodexes (Code of Practice, CoP) in verschiedenen Arbeitsgruppen beteiligt, darunter Lobbyisten nahezu aller relevanten Anbieter von GPAI-Modellen. Gemäß Art. 56 KI-Verordnung hätte die finale Version des Kodexes am 2. Mai veröffentlicht sein müssen. Schließlich wurde der 10. Juli daraus.

„Rigorose Analysen“ gefordert

Im CoP verpflichten sich die Unterzeichner, eine Modelldokumentation zu erstellen, sie auf dem neuesten Stand zu halten und relevante Informationen mit nachgelagerten AI-Systemintegratoren und dem AI-Büro der EU auf Anfrage zu teilen. Der Kodex soll das durch die Bereitstellung eines standardisierten Modelldokumentationsformulars erleichtern. Die Anbieter seien außerdem dafür verantwortlich, die Qualität, Sicherheit und Integrität der von ihnen dokumentierten Informationen zu gewährleisten.

Für GPAI-Modelle mit systemischem Risiko legt der Kodex einen Rahmen fest, der die Verfahren des Anbieters zur Risikobewertung, Risikominderung und Steuerung detailliert beschreibt. Die Unterzeichner des Kodexes verpflichten sich, technische Sicherheitsvorkehrungen gegen Probleme zu treffen, die während des gesamten Lebenszyklus des Modells auftreten können.

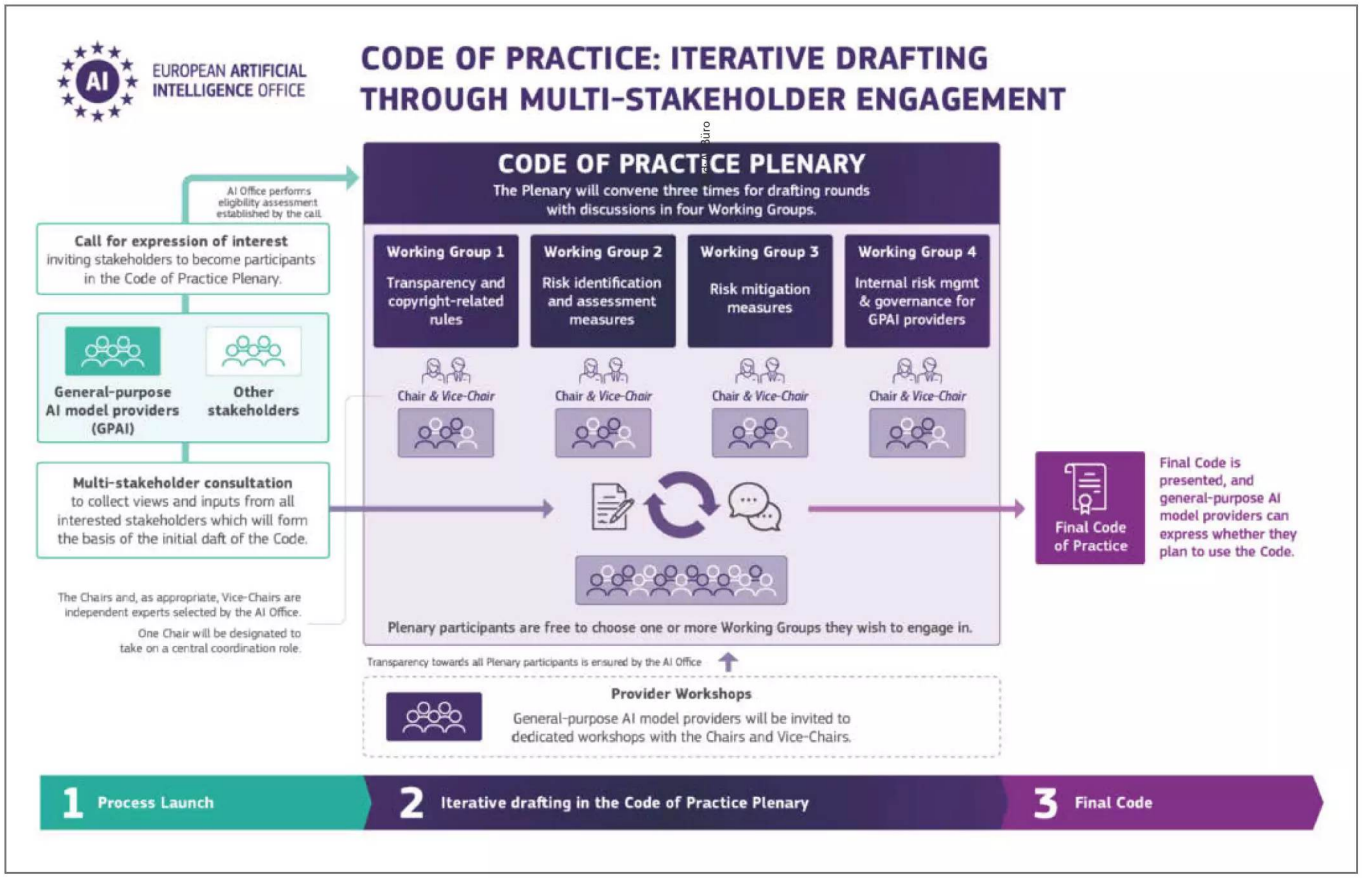
Der Kodex legt großen Wert auf die in der KI-Verordnung etablierte Rechenschaftspflicht. Er fordert regelmäßige Sicherheitsmodellberichte, die die Einhaltung des Kodexes dokumentieren. Ein weiterer wichtiger Punkt, den der Kodex vorschreibt, ist die Notwendigkeit externer, unabhängiger Audits aller Systemrisikomodelle. Außerdem müssen Anbieter solcher Modelle demnach Kanäle für die Meldung von Vorfällen und Whistleblowing einrichten.

Streitpunkt Urheberrecht

Besonders strittig waren die Bestimmungen zum Urheberrecht. Wie erwähnt sind die Anbieter von GPAI-Modellen verpflichtet, eine Strategie zur Einhaltung des EU-Urheberrechts zu entwickeln (Art. 53 KI-Verordnung). Hier soll der Kodex eine Regelungslücke schließen, denn wie eine solche Strategie aussehen könnte, verrät die KI-Verordnung selbst nicht.

Dem CoP zufolge dürfen Modellanbieter das Web mit Crawlern abgrasen und müssen sich lediglich an Paywalls sowie an ausgesprochene Opt-outs halten. Diese Opt-out-Erklärungen auf Websites müssen maschinenlesbar sein – wie es das deutsche Urheberrecht ohnehin vorsieht. Alle Unterzeichner versprechen, sich an das robots.txt-Protokoll zu halten. Auch andere maschinenlesbare Opt-outs in den Metadaten, „die als Industriestandard etabliert sind“, müssen sie berücksichtigen. Die Unterzeichner verpflichten sich demnach auch, Rechteinhaber darüber zu informieren, welche Crawler sie einsetzen und wie diese mit robots.txt-Dateien umgehen.

Modellanbieter müssen „angemessene Anstrengungen“ aufwenden, um Trainingsmaterial urheberrechtlich zu prüfen. Mit der Formulierung der an-



Mit einem Organigramm erläutert das AI-Büro, wie der Entwicklungsprozess zum fertigen Verhaltenskodex abgelaufen ist.

gemessenen Anstrengungen hat man auch andere Verpflichtungen relativiert: Während im zweiten Entwurf des Kodexes die Ausgabe von zu urheberrechtlich geschützten Werken sehr ähnlichem Output noch explizit verboten war, sollen Anbieter nun dieses Risiko einer Rechtsverletzung nur noch möglichst minimieren. Gerade an dieser Stelle scheint die KI-Anbieter-Lobby ganze Verwässerungsarbeit geleistet zu haben.

Entsprechend erzürnt reagierten Interessenverbände der Rechteinhaber. Das European Writers' Council (EWC) beispielsweise war bei der Ausarbeitung des Kodexes selbst am Tisch und zeigte sich nach der Verabschiedung entsetzt. Kritik hagelte es auch aus der Wissenschaft. In einem Beitrag für das Verfassungsblog stellte Martin Ebers, Vorsitzender der Robotics & AI Law Society (RAILS), die Frage, ob der Verhaltenskodex von der KI-Lobby momentan als trojanisches Pferd zur Neugestaltung der KI-Verordnung genutzt werde.

Unbenommen der Bedenken aus den USA bleibt aber Tatsache, dass der GPAI Code of Practice Anfang August gesetzesähnliche Bedeutung erlangt hat. Anbietern, aber auch Betreibern, bleibt keine Wahl: Sie müssen sich dieses Dokument sehr genau ansehen und die enthaltenen Compliance-Forderungen umsetzen, wenn sie Sanktionen vermeiden wollen. Es wird viel Papier produziert werden müssen, was Anwaltskanzleien und Beratungsfirmen goldene Zeiten bescheren dürfte. (hob) c't

Vorschau: c't Apple-Einkaufsratgeber

Ab dem 21.11.25 im Handel und auf ct.de

Tests zu iPhone, iPad, Mac & mehr

Wer viel Geld für ein Apple-Produkt in die Hand nimmt, möchte gut beraten sein: Welches Modell ist die richtige Wahl? Wo liegen die Unterschiede? Was braucht man, was nicht? Die Redakteure des c't-Schwestermagazins Mac & i klären auf, das nächste c't-Sonderheft bündelt die Tests zu den neuesten iPhones von 2025, bespricht das iPad- und Mac-Portfolio und berät, wie und mit welcher Hardware der Umstieg auf den Mac gelingt. Zudem steht Zubehör auf dem Prüfstand, von der Apple Watch bis hin zu Kopfhörern.

Weitere c't-Sonderhefte: heise.de/s/00MxL

Themenschwerpunkte

Die aktuellen iPhones & iPads

- iPhone 17, 17 Pro & Air – welches Modell sich lohnt
- iPhone 16e – das günstigste iPhone von 2025
- iPad 11 – reicht das Einsteigertablet?
- iPad Air M3 – mit viel Pro-Power
- iPad Pro – Apples Spitzenmodell

Mac-Rechner im Testlabor

- Hardware-Beratung für den Umstieg auf macOS
- MacBook Air M4 – elegant, günstig, aber auch gut?
- MacBook Pro M4 – Leistung dank Pro- und Max-Chips
- Mac mini M4 Pro und iMac
- M4 – für den Schreibtisch
- Mac Studio M3 Ultra – der kompakte Profi-Rechner

Zubehör auf dem Prüfstand

- Kleine externe Festplatten ab 4 TByte im Vergleich
- Ultrabreite Displays für den Anschluss am Mac
- Apple Watch 11, Ultra 3 und SE 3 – wer braucht welche?
- AirPods Pro 3 und AirPods 4 mit Geräuschunterdrückung
- Over-Ear-Kopfhörer im großen Vergleich

WIR TEILEN KEIN HALBWISSEN.
WIR SCHAFFEN FACHWISSEN.

Workshop 13. November

Einführung in GitLab

Erfahren Sie, wie Sie GitLab einrichten, konfigurieren und anpassen. Außerdem lernen Sie, wie Sie eine eigene Instanz der Entwicklungsplattform betreiben.

Workshop 26. – 27. November

Einführung in den Kea DHCP Server

Erfahren Sie alles über Kea-DHCP-Software auf Unix- und Linux-Systemen. Sie lernen mehr über die Installation, Konfiguration und Betrieb des Systems.

Webinar 27. November

MCP verstehen: So arbeiten KI-Agenten für Sie

Wir stellen vor, was es mit dem Model Context Protocol auf sich hat, wie man sie im Alltag einsetzt und welche neuen und altbekannten Sicherheitslücken MCP aufreißt.

Workshop 2. – 3. Dezember

Docker und Container in der Praxis

Nach dem Workshop sind Sie in der Lage, eine eigene Infrastruktur in Betrieb zu nehmen oder eigene Docker-Abbilder zu verpacken.

Mehr anzeigen ▲

heise.de/ct/Events

SUPPORT ME

🙏 Hope my post useful for you, if you want support me please following one of the ways:

Buy or Renew Premium Account

👉 Rapidgator: https://rapidgator.net/account/registration/ref/49023

👉 Nitroflare: https://nitroflare.com/payment?webmaster=194862

⚠ Note: Please DON'T turn on VPN when making payment.

Donate Directly

USDT (TRC20):

TFniVipHpFsPVrUHBLsvkZJV4Mjj1MUz96

DOGE (Doge Network):

DCfVVnvNaVtxQbWyfpWsihbGnvpkuYdtJS

Every little support helps me to keep going and create more content.

THANK YOU SO MUCH!