KI-4-Everyone · Daily News

7. Mai 2026

SAFE

ChatGPT kann jetzt Vertrauenspersonen bei Krisen alarmieren

OpenAI führt eine optionale Funktion ein: Erwachsene können einen Notfallkontakt hinterlegen, der bei erkannten Selbstgefährdungshinweisen benachrichtigt wird.

MARKT

US-Energieminister und NVIDIA: KI soll ihren eigenen Strom sichern

Chris Wright und NVIDIAs Ian Buck argumentieren, dass KI aktiv beim Aufbau der Energieinfrastruktur helfen wird, die sie selbst benötigt.

SAFETY

ChatGPT bekommt einen Notfallkontakt: OpenAI fuehrt 'Trusted Contact' ein

Wer ChatGPT nutzt, kann kuenftig eine Vertrauensperson hinterlegen, die alarmiert wird, wenn der Chatbot Hinweise auf Selbstgefaehrdung erkennt.

Ein Chatbot, der bei ernsten Sorgen um das Leben seines Nutzers nicht mehr nur Hotline-Nummern einblendet, sondern aktiv einen Menschen aus dem persoenlichen Umfeld benachrichtigt: Mit dieser Idee geht OpenAI einen Schritt, der die Rolle von ChatGPT spuerbar verschiebt. Das Werkzeug wird damit erstmals Teil eines persoenlichen Sicherheitsnetzes - mit allen Chancen und allen heiklen Fragen, die das aufwirft. Die Funktion heisst 'Trusted Contact' und ist optional.

OpenAI kuendigte das Feature an: Erwachsene Nutzerinnen und Nutzer koennen eine Person ihres Vertrauens als Notfallkontakt in ChatGPT hinterlegen - etwa Freunde, Familienmitglieder oder Betreuende. Erkennt das System in einem Gespraech ernste Hinweise auf Selbstverletzung oder suizidale Inhalte, wird dieser hinterlegte Kontakt benachrichtigt. Die Funktion ist als zusaetzliche Sicherheitsebene angelegt und nicht standardmaessig aktiv, sondern muss von der Nutzerin oder dem Nutzer selbst eingerichtet werden. Berichte ueber den Start kamen am 7. Mai aus dem OpenAI-Blog sowie von The Verge.

Der Schritt kommt nicht aus dem Nichts. KI-Chatbots sind in den vergangenen Monaten zunehmend in die Kritik geraten, weil Menschen in psychischen Krisen lange Gespraeche mit ihnen fuehren - manchmal anstelle eines Gespraechs mit einem Menschen. Bisher reagierten Systeme wie ChatGPT in solchen Faellen vor allem mit Standardhinweisen auf professionelle Hilfsangebote. Mit 'Trusted Contact' bezieht OpenAI erstmals eine reale Bezugsperson aus dem Leben des Nutzers ein. Das verschiebt die Verantwortung: weg vom alleinigen Dialog Mensch-Maschine, hin zu einem Dreieck aus Nutzer, KI und Vertrauensperson. Fuer Angehoerige kann das eine Entlastung sein - oder eine Ueberforderung, je nachdem, wie ein Alarm sich konkret anfuehlt.

Im Material bleibt vieles offen. Unklar ist, wie genau OpenAI 'serious self-harm concerns' technisch erkennt, wie oft das System anschlaegt und wie hoch die Rate falscher Alarme sein koennte - ein Chatbot, der Kontakte alarmiert, obwohl keine echte Krise vorliegt, waere ein erhebliches Vertrauensproblem. Ebenfalls nicht im Material belegt ist, ob die Funktion weltweit oder nur in bestimmten Laendern startet, wie der Notfallkontakt informiert wird (per E-Mail, SMS, App-Nachricht?) und ob die hinterlegte Person dem Verfahren vorab zustimmen muss. Heikel ist zudem der Datenschutz: Eine Vertrauensperson bekaeme zumindest indirekt Hinweise auf den seelischen Zustand eines Erwachsenen - mit allen rechtlichen Fragen, die das in Europa aufwirft.

Beobachten lohnt sich in den naechsten Tagen vor allem, ob OpenAI nachliefert: Wie sieht der Alarm konkret aus, was genau erfaehrt der Notfallkontakt, und gibt es Reaktionen von Fachleuten aus Psychiatrie und Suizidpraevention? Auch andere Anbieter werden vermutlich Position beziehen muessen - denn wenn ChatGPT einen Notfallkontakt einfuehrt, wird die Frage, warum konkurrierende Chatbots das nicht tun, schnell unangenehm.

MARKT

Microsoft: Nur 17,8 % der Weltbevölkerung nutzen KI aktiv

Microsoft hat seinen Global AI Diffusion Report veröffentlicht. Demnach nutzen weltweit 17,8 % der Menschen KI. Der Bericht zeigt eine wachsende Kluft zwischen globalem Norden und Süden.

MARKT

KI-Boom lässt Mainboard-Verkäufe einbrechen

Der Mainboard-Markt erlebt einen drastischen Einbruch. Der Grund: KI-Workloads treiben eine beispiellose Knappheit bei Komponenten. Wer ein normales Mainboard kaufen will, findet immer weniger Auswahl.

MARKT

Anthropic erhöht Claude-Nutzungslimits und schließt Deal mit SpaceX

Anthropic gibt Claude-Nutzern höhere Nutzungslimits. Gleichzeitig vereinbarte das Unternehmen einen Compute-Deal mit SpaceX. Details zur Größe des Deals nennt Anthropic nicht.

SAFE

KI-generierter Masseninhalt schadet Online-Communities

Automatisch erzeugter KI-Inhalt überflutet Foren und soziale Plattformen. Communities verlieren dadurch Qualität und echten Austausch. Das Problem gilt inzwischen als ernstes Risiko für den offenen Diskurs im Netz.

PROD

GitHub erklärt, wie du KI-generierte Pull Requests richtig prüfst

KI-Agenten erstellen heute massenhaft Pull Requests in Codeprojekten. GitHub hat einen praktischen Leitfaden veröffentlicht: Worauf du achten musst, wo Fehler stecken und wie du technische Schulden früh erkennst.

OS

LLM-Training beschleunigen mit Unsloth und NVIDIA

Unsloth verspricht in Kombination mit NVIDIA-Hardware deutlich schnelleres Training von Sprachmodellen. Für alle, die eigene Modelle trainieren, könnte das Kosten und Zeit sparen. Genaue Speedup-Zahlen sind im Material nicht angegeben.

PROD

Google stellt Fitbit Air vor: Fitness-Tracker ohne Display für 100 Dollar

Google präsentiert den Fitbit Air – einen bildschirmlosen Fitness-Tracker. Das Gerät kostet 100 Dollar und ist ab sofort vorbestellbar. Google ersetzt damit die bisherige Fitbit-App durch eine neue Google-Health-App.

PROD

OpenAI bringt neue Sprachmodelle für Echtzeit-Voice in die API

OpenAI stellt neue Realtime-Voice-Modelle in seiner API vor. Sie sollen Sprache übersetzen, transkribieren und dabei natürlicher klingen als bisher. Entwickler können die Modelle direkt über die OpenAI-API nutzen.

TOOLS & HANDS-ON

OS

Googles Gemma 4 versteht Bilder und Text – und läuft auf deinem Rechner

Gemma-4-26B-A4B-it ist ein Open-Source-Modell, das nur 4 von 26 Milliarden Parametern gleichzeitig aktiviert – dadurch antwortet es schneller und braucht weniger Rechenleistung. Es verarbeitet sowohl Text als auch Bilder.

PROD

ChatGPT testet Werbung – mit Versprechen zu Datenschutz und Transparenz

OpenAI schaltet erste Anzeigen in ChatGPT, um den kostenlosen Zugang zu finanzieren. Werbung soll klar gekennzeichnet sein und Antworten nicht beeinflussen.

PROD

Parloa baut Kundenservice-Agenten auf OpenAI-Basis

Parloa nutzt OpenAI-Modelle, um sprachgesteuerte Kundenservice-Agenten für Unternehmen bereitzustellen. Die Agenten lassen sich entwerfen, simulieren und in Echtzeit einsetzen.

PROD

KI-Podcasts direkt in Spotify speichern – neues Tool macht es möglich

"Save to Spotify" ist ein Kommandozeilen-Tool für KI-Agenten wie Claude Code oder OpenAI Codex. Wer Recherchen per KI in Audio-Zusammenfassungen umwandelt, kann diese damit direkt in Spotify ablegen.

PROD

GitHub erklärt, wie es seinen Copilot-Agenten auf Fehler testet

GitHub beschreibt, wie man KI-Agenten überprüft, wenn es keine eindeutig richtige Antwort gibt. Statt starrer Skripte nutzt GitHub Copilot eine Methode namens "dominatory analysis" für seinen Coding-Agenten.

PROD

GeForce NOW: Gaijin-Spiele jetzt mit Single Sign-On zugänglich

NVIDIAs Cloud-Gaming-Dienst GeForce NOW unterstützt ab sofort das Single Sign-On von Gaijin. Wer Gaijin-Spiele nutzt, muss sich beim Start nicht mehr separat einloggen.

AUSBLICK

Keine Termine gemeldet.