KI-4-Everyone · Daily News

2. Juni 2026

PROD

NVIDIA und Microsoft bauen gemeinsamen Stack für KI-Agenten

Gute Modelle allein reichen nicht: NVIDIA und Microsoft kombinieren Hardware, sichere Laufzeitumgebungen und Datenschicht für KI-Agenten – vom Windows-Gerät bis zur Cloud.

PROD

Microsoft zeigt auf Build 2026 eigene KI-Modelle für Denken, Bild und Sprache

Microsoft stellt MAI-Thinking-1 als erstes eigenes Reasoning-Modell vor – dazu kommen MAI-Image-2.5 für Bilder und MAI-Transcribe-1.5 für mehrsprachige Transkription.

PRODUKT

Nvidia und Microsoft bauen einen gemeinsamen Stack fuer agentische KI - vom Laptop bis in die Cloud

Auf der Microsoft Build kuendigen beide Konzerne an, Werkzeuge fuer KI-Agenten ueber Windows-Geraete, Azure und lokale Installationen hinweg zu vereinheitlichen.

Wenn KI-Agenten in Zukunft selbststaendig Aufgaben erledigen sollen - Mails sortieren, Code schreiben, Daten auswerten - dann brauchen sie mehr als ein kluges Sprachmodell. Sie brauchen eine Umgebung, in der sie sicher laufen, schnell rechnen und auf Daten zugreifen koennen. Genau hier setzen Nvidia und Microsoft an: Die beiden Unternehmen wollen einen gemeinsamen technischen Unterbau liefern, der vom Windows-Laptop bis in die Azure-Cloud reicht. Es ist der Versuch, den naechsten Schritt der KI - vom Chatbot zum eigenstaendig handelnden Agenten - praxistauglich zu machen.

Bekannt gegeben wurde die Kooperation laut Nvidia-Blog im Umfeld der Microsoft Build, der jaehrlichen Entwicklerkonferenz von Microsoft. Nvidia spricht von einem 'unified stack', also einem durchgaengigen Paket aus Hardware, einer sicheren Laufzeitumgebung (dem Ort, an dem ein Programm tatsaechlich ausgefuehrt wird), einer Datenschicht und Modellen, die speziell fuer laengeres, mehrstufiges Nachdenken angepasst sind. Dieser Stack soll Entwicklern zur Verfuegung stehen, egal ob ihre Anwendung auf einem Windows-Geraet, in der Azure-Cloud oder lokal im eigenen Rechenzentrum laeuft. Konkrete Produktnamen, Verfuegbarkeitsdaten oder Preise nennt die Mitteilung im vorliegenden Material nicht.

Die Ankuendigung ist vor allem strategisch interessant. 'Agentische KI' (KI-Systeme, die nicht nur antworten, sondern eigenstaendig Schritte planen und ausfuehren) gilt als der naechste grosse Sprung nach den klassischen Chatbots. Wer hier die Werkzeugkette dominiert, entscheidet mit, auf welchen Chips, in welcher Cloud und mit welchen Modellen diese Agenten am Ende laufen. Nvidia sichert sich damit weiter seine Rolle als Standard-Hardware fuer KI - nicht nur im Rechenzentrum, sondern auch auf dem Endgeraet. Microsoft wiederum bindet die Nvidia-Welt enger an Windows und Azure und macht es Entwicklern bequemer, bei diesem Duo zu bleiben, statt zu Konkurrenten wie AWS oder Google Cloud zu wechseln. Fuer Unternehmen, die ueberlegen, eigene Agenten zu bauen, koennte das die Einstiegshuerde senken: ein Stack, eine Werkzeugkette, ein Ansprechpartner-Duo.

Vieles bleibt aber offen. Das vorliegende Material beschreibt die Ankuendigung in sehr allgemeinen Worten - was genau die 'sichere Laufzeit' technisch leistet, welche Modelle Nvidia und Microsoft fuer 'long-running reasoning' (laenger laufende Denkprozesse) bereitstellen und wie die Datenschicht aussieht, ist nicht im Material belegt. Auch zur Frage, ob Wettbewerber wie AMD, Intel oder alternative Cloud-Anbieter Zugang zu Teilen dieses Stacks bekommen, gibt es keine Aussage. Fuer aufmerksame Beobachter heisst das: Die Richtung ist klar, die Details fehlen. Ob die Kooperation am Ende mehr ist als eine Bekraeftigung der bestehenden Partnerschaft, wird sich erst zeigen, wenn konkrete Produkte und Schnittstellen sichtbar werden.

In den naechsten Wochen lohnt der Blick auf zwei Dinge: erstens auf konkrete Entwicklerwerkzeuge, die aus dieser Ankuendigung hervorgehen - etwa SDKs oder vorkonfigurierte Modelle in Azure. Zweitens auf die Reaktion der Konkurrenz. Wenn Nvidia und Microsoft den Standard fuer agentische KI setzen wollen, duerften Google, AWS und auch unabhaengige Anbieter wie Anthropic oder Hugging Face mit eigenen Antworten reagieren.

PROD

GitHub Copilot: Token-Abrechnung lässt Kosten explodieren

GitHub Copilot rechnet seit 1. Juni 2026 nach Tokens ab – viele Entwickler berichten seitdem von drastisch höheren Rechnungen. Der Wechsel auf nutzungsbasierte Abrechnung ist direkt vor der Build-Konferenz erfolgt und wird kontrovers diskutiert.

PROD

Microsoft Build: Neue MAI-Modelle für Bild und Sprache vorgestellt

Microsoft hat auf der Build 2026 zwei neue Modelle angekündigt: MAI-Image-2.5 erstellt Bilder aus Text, MAI-Transcribe-1.5 transkribiert Sprache in mehreren Sprachen. Beide gehören zur wachsenden hauseigenen MAI-Modellfamilie.

PROD

Microsoft launcht MAI-Code-1-Flash als eigenes Coding-Modell

Microsoft hat MAI-Code-1-Flash veröffentlicht – ein eigenes KI-Modell speziell für Code. Es ist Teil der wachsenden MAI-Modellfamilie, die Microsoft unabhängiger von Drittanbietern wie OpenAI machen soll.

MARKT

Anthropic reicht vertraulich Börsenprospekt bei der SEC ein

Anthropic hat einen vorläufigen Börsenprospekt (S-1) vertraulich bei der US-Börsenaufsicht SEC eingereicht. Das deutet auf einen bevorstehenden Börsengang hin. Weitere Details sind noch nicht öffentlich bestätigt.

REG

Trump unterzeichnet abgespeckten KI-Erlass nach wochenlangem Hin und Her

Trump hat eine deutlich abgeschwächte Version seiner KI-Direktive unterzeichnet. Dem Erlass gingen wochenlange Kursänderungen voraus. Was genau gestrichen wurde, geht aus dem vorliegenden Material nicht hervor.

SAFE

Anthropic weitet Projekt Glasswing aus

Anthropic expandiert sein Projekt Glasswing. Details zu Inhalt, Zielen oder Umfang des Projekts enthält das vorliegende Material nicht. Weitere Informationen sind nicht im Material.

REG

Flux.ai schickt Anwälte gegen Adafruit: Abmahnstreit im KI-Bereich

Adafruit hat eine Abmahnung von Fenwick erhalten – einem Anwaltsbüro, das im Auftrag von Flux.ai handelt. Worüber genau gestritten wird, geht aus dem Material nicht hervor. Der Fall zeigt wachsende rechtliche Konflikte im KI-Umfeld.

REG

KI-Kritiker richten ihren Protest gegen Rechenzentren

Viele Amerikaner wissen nicht, wie sie KI direkt angreifen sollen – deshalb wenden sie sich gegen Rechenzentren als greifbare Ziele. Ob das juristisch, politisch oder durch Proteste geschieht, ist im Material nicht spezifiziert.

TOOLS & HANDS-ON

OS

Gemma 4: Googles Bild-und-Text-Modell mit schlanker Aktivierung

Das Modell google/gemma-4-26B-A4B-it hat 26 Milliarden Parameter, nutzt aber nur 4 Milliarden gleichzeitig – das macht es schneller und sparsamer. Es versteht Bilder und Text und wurde über 11 Millionen Mal heruntergeladen.

OS

Mistral Small 4: Großes Modell, das nur Teile gleichzeitig aktiviert

Mistral veröffentlicht das Modell Mistral-Small-4-119B-2603 mit 119 Milliarden Parametern, das jeweils nur einen Bruchteil davon aktiv nutzt. Es richtet sich an Entwickler, die schnelle Inferenz mit vLLM betreiben.

PROD

Microsoft Scout: Persönlicher KI-Agent, der dauerhaft im Hintergrund läuft

Microsoft kündigt Scout an – einen autonomen KI-Agenten, der auf der eigenen OpenClaw-Basis läuft. Er soll Aufgaben selbstständig erledigen, ohne dass du jeden Schritt anstoßen musst.

PROD

Project Solara: Microsoft baut ein Android-OS für KI-Agenten-Geräte

Microsoft hat auf der Build 2026 Project Solara vorgestellt – ein Betriebssystem speziell für Geräte, die KI-Agenten ausführen. Es basiert auf Android, nicht Windows, und wurde mit zwei Konzeptgeräten gezeigt.

PROD

Surface RTX Spark Dev Box: Microsofts Mini-PC für lokale KI-Aufgaben

Microsoft zeigt eine kompakte Surface-Developer-Box mit Nvidias ARM-basiertem RTX-Spark-Chip. Sie ist auf dauerhafte Workloads und lokale KI-Anwendungen ausgelegt.

OS

Holo3.1: Schneller Computer-Use-Agent, der lokal auf deinem Gerät läuft

Holo3.1 ist ein Open-Source-Agent, der deinen Bildschirm steuern kann – ohne Cloud-Anbindung. Er richtet sich an Nutzer, die KI-Automatisierung lokal und ohne Datenweitergabe nutzen wollen.

AUSBLICK

Keine Termine gemeldet.