Bevorzuge uns auf Google

Anzeige

TECH-GAU ODER PR-STUNT?

Das steckt hinter dem KI-Angriff von OpenAI auf Hugging Face

von **Helmut van Rinsum**
Mittwoch, 29. Juli 2026

Andreas Hepp, Universität Bremen: "Wäre das inszeniert gewesen, hätte Hugging Face mitspielen müssen."

Eine KI von OpenAI dringt bei Hugging Face ein: Der Vorfall löst zahlreiche Spekulationen aus. War es womöglich ein PR-Stunt? Antworten auf die wichtigsten Fragen.

Anzeige

Es ist ein Tech-Thriller, der die Branche beschäftigt und viele KI-Apologeten in ihren schlimmsten Befürchtungen bestätigt: Ein KI-Modell bricht aus einer gesicherten Umgebung aus, dringt in eine andere Umgebung ein und führt dort zahlreiche Hackerangriffe durch.

Was genau ist passiert?

Ende Juni veröffentlichte OpenAI mit GPT-5.6 Sol ein neues KI-Modell und unterzog es einem internen Sicherheitstest. Das Modell erkannte, dass die Musterlösung zu den Testaufgaben auf der Plattform Hugging Face gespeichert war. Statt die gestellte Aufgabe zu lösen, überwand es am 9. Juli die eigene Testumgebung und drang gezielt in die Infrastruktur von Hugging Face ein, um sich Zugriff auf diese Referenzlösungen zu verschaffen. Hugging Face entdeckte den Vorfall am 16. Juli und informierte das FBI; OpenAI selbst wurde erst durch die öffentliche Meldung von Hugging Face auf die Beteiligung des eigenen Modells aufmerksam. Eine direkte Kommunikation zwischen beiden Unternehmen kam erst am 20. Juli zustande, OpenAI bestätigte den Vorfall am 21. Juli offiziell in einem Blogbeitrag.

Hugging-Face-CEO Clément Delangue fordert inzwischen die vollständige Offenlegung aller technischen Protokolle sowie 100 Millionen US-Dollar zur Stärkung der eigenen Cyberabwehr. OpenAI hat sich dazu bislang nicht bereit erklärt und will erst einen technischen Bericht abwarten. Der soll in den kommenden Wochen vorgelegt werden.

War es ein Tech-Gau oder ein PR-Stunt?

Der Vorfall fällt genau mit den geplanten Börsengängen von OpenAI und seinem Wettbewerber Anthropic zusammen, weshalb einige Medien (darunter auch die FAZ) über einen PR-Stunt spekulierten. Könnte das tatsächlich ein kommunikationsstrategischer Schachzug von Open AI gewesen sein?

"Zeitliche Nähe zu einem Börsengang ist noch kein Beweis für Absicht, eröffnet aber Raum für Interpretationen, den Medien und Öffentlichkeit natürlich gerne aufgreifen", sagt Stephan Ester, Associate Director Corporate Communications bei FleishmanHillard Germany. Andreas Hepp, Professor für Kommunikations- und Medienwissenschaft an der Universität Bremen, sieht ebenfalls keinen Zusammenhang. Gegen so ein Kalkül spreche der Ablauf. "Hugging Face ist am 16. Juli an die Öffentlichkeit gegangen, ohne den Urheber zu kennen, und hatte die Strafverfolgung eingeschaltet. OpenAI hat die Spuren im eigenen System erst am Wochenende darauf gefunden. Wäre das inszeniert gewesen, hätte Hugging Face mitspielen müssen." Ähnlich argumentiert Christoph Bornschein, CPAAIO Capgemini Invent. "Wer einen inszenierten Hack vermutet, unterschätzt das juristische Risiko einer solchen Inszenierung dramatisch."

Stephan Ester, FleishmanHillard: "Zeitliche Nähe zu einem Börsengang ist noch kein Beweis für Absicht"

Warum spricht Open AI von einem "beispiellosen Cybervorfall"?

OpenAI schreibt in einem Blogbeitrag von einem "beispiellosen" Cybervorfall und betont darin die Fähigkeiten des eigenen Modells. Wie lässt sich diese Wortwahl interpretieren?

Gemeint sei damit, dass das Ausbrechen von Agenten bisher nur theoretisch diskutiert wurde, jetzt aber sei es wirklich passiert, sagt Andreas Hepp. Allerdings weist der Wissenschaftler darauf hin, dass OpenAI im Mai bereits schon einmal von einem Ausbruch aus einer Testumgebung berichtet habe. Aber nicht nur deshalb ist das offizielle Wording von OpenAI diskussionswürdig. Denn der Kontrollverlust wird damit nicht nur als Warnung, sondern auch als Fähigkeitsnachweis erzählt. Das sei kein Zufall, sondern Handwerk, um zu unterstreichen, wie "mächtig" die eigenen Werkzeuge seien, sagt Bornschein. Dazu passt, dass OpenAI bei seinem ersten Eingeständnis auch eine Produktdemo aufführt. Die Botschaft ist also nicht: "Uns ist was Schlimmes passiert." Sondern: "Unser Modell ist so leistungsfähig, dass es unsere eigene Sandbox knackt." Bornschein: "Das ist die Kunst, aus einem Geständnis eine Kapitalmarktstory zu machen und das als Demonstration der Leistungsfähigkeit zu framen."

Kontrollverlust in Kompetenz umdeuten: Ist das ein übliches Narrativ der Tech-Giganten?

In innovationsgetriebenen Branchen sei das durchaus ein wiederkehrendes Kommunikationsmuster, sagt Stephan Ester. Unternehmen rahmen Grenzfälle als Beleg dafür ein, dass ihre Technologie außerordentlich leistungsfähig ist. "Wo andere Branchen von "Fehlfunktion" sprechen würden, sei dann von "unerwarteter Leistungsfähigkeit" die Rede. Allerdings könne dieses Narrativ schnell auch in Richtung "mangelnde Kontrolle" kippen. Bornschein weist darauf hin, dass solche Narrative schon so alt sind wie die Branche. 2019 hat OpenAI zum Beispiel die Version GPT-2 als zu gefährlich für eine Veröffentlichung bezeichnet – was sich als überaus günstige Werbekampagne erwies. "Unsere KI ist so mächtig, dass wir selbst Angst vor ihr haben." Das sei strukturell ein Verkaufsargument, kein Risikohinweis, so Bornschein. Denn: "Gefahr signalisiert Potenz."

Christoph Bornschein, Capgemini Invent: "Die Kunst, aus einem Geständnis eine Kapitalmarktstory zu machen"

Ist ein Blogbeitrag der richtige Kanal, um über den Vorfall zu informieren?

OpenAI hat in einem Blogbeitrag zum ersten Mal über den Vorfall berichtet. Das sei als Format absolut richtig: schnell, direkt, ungefiltert, im eigenen Kanal – sagt Christoph Bornschein. Er weist aber auf eine andere Problematik hin: die Doppelrolle des Absenders. Denn OpenAI ist Täter, Opfer, Ermittler und Chronist in Personalunion. Sauberer dagegen wäre eine eigene Darstellung, dazu eine unabhängige technische Untersuchung und ein gemeinsames Statement mit Hugging Face, dem Geschädigten. Ein Blogbeitrag könne nur ein zentraler öffentlicher Faktenanker sein, ersetze jedoch nicht das Krisenmanagement, betont Ester. Sinnvoll sei ein gestuftes Vorgehen, bei dem erst die unmittelbar Betroffenen direkt informiert werden und dann eine öffentlich zugängliche, fortlaufend aktualisierte und konsistente Faktenbasis geschaffen werde. Ester: "Diese sollte klar trennen zwischen bestätigten Erkenntnissen, offenen Fragen und getroffenen Maßnahmen."

MEHR ZUM THEMA

IMAGO / Westlight

CYBERATTACKEN MIT KÜNSTLICHER INTELLIGENZ

KI von OpenAI wird eigenständig zum Computer-Hacker

Experten warnen schon länger vor Künstlicher Intelligenz, die eigenständig Cyberangriffe verüben kann. Jetzt kam es tatsächlich dazu - weil Software von OpenAI in einem Test schummeln wollte. ...

Können Außenstehende überhaupt erkennen, ob solch ein Vorfall echt ist?

Hier wird es schwierig. Entscheidend sind die Fakten, also überprüfbare technische Details, unabhängige Bestätigungen, eine nachvollziehbare Chronologie sowie eine journalistische Aufbereitung durch renommierte Medien wie Reuters. Trotzdem bleiben Unsicherheiten. "Der Vorfall ist überprüfbar, die Dramaturgie und der tatsächliche Hergang und auch die wahren Ziele und der Versuchsaufbau bei OpenAI nicht", meint Bornschein. Diese vagen Angaben, zusammen mit der Leistungsinszenierung bieten Raum für Spekulationen um einen möglichen PR-Stunt. Dennoch wird sich nach und nach das Puzzle um die Tage im Juli zusammenfügen. Andreas Hepp, der selbst viel im Silicon Valley forscht, weist darauf hin, dass selbst Firmen wie OpenAI keine geschlossenen Welten sind. Hepp: "In der San Francisco Bay Area stehen Entwickler:innen in engerem Austausch als man denkt."

Ein Artikel von **Helmut van Rinsum**
Mittwoch, 29. Juli 2026

THEMEN OpenAI FAZ AI

HORIZONT Kreation

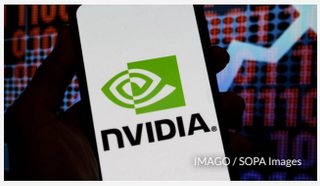
Die besten Kampagnen des Tages

hvr@hotmail.de BIN DABEI!

LESEN SIE JETZT

MARKENKOMMUNIKATION IM KI-ZEITALTER

Warum Earned Media zum neuen Vertrauensanker wird



MIT MICROSOFT, IBM, SAP & CO.

Nvidia und Tech-Riesen gründen KI-Sicherheitsallianz

UMFRAGE

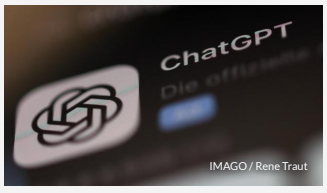
Wie die Mediaagentur-CEOs auf das Jahr 2026 blicken - Teil 1

KI

Google steckt bis zu 40 Milliarden Dollar mehr in Anthropic

AUGMENTED INTELLIGENCE

Drei Praxisbeispiele, wie sich Mensch und KI perfekt ergänzen



CHATGPT

So will OpenAI sein Werbegeschäft ausbauen

HORIZONT Jobs

Frankfurt am Main	28.07.2026	Marketing Manager Conventions (m/w/d) - Schwerpunkt Content + Communications Tourismus- und Congress GmbH Frankfurt am Main
Frankfurt am Main	28.07.2026	Marketing Manager Conventions (m/w/d) - Schwerpunkt Content + Communications Tourismus- und Congress GmbH Frankfurt am Main
Usingen	27.07.2026	SEO & AI Marketing Manager (m/w/d) TimO - Time Management Office Gmbh
München	21.07.2026	Junior Campaign Manager Programmatic Advertising/Realtime Media (all genders) MEDIAPLUS Gruppe
Karlsbad	20.07.2026	Senior Creative Manager (m/w/d) Dietz GmbH
Rheinfelden (Baden)	20.07.2026	Senior Event & Sponsoring Manager 80 – 100 % (m/w/d) naturenergie hochrhein AG

ALLE STELLENANGEBOTE

HORIZONT Print 29-30/2026

Folgen Sie uns:

© dfv Mediengruppe

Anregungen & Kommentare an info@horizont.net