

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

*Hanna Muralt Müller 11.04.2026*

**Neuer KI-Agent kann Stromnetze, Spitäler oder Militäranlagen lahmlegen. Bereits hat er kritische Sicherheitslücken aufgedeckt.**

*Red. Als Vizekanzlerin im Bundeshaus von 1991 bis 2005 leitete die Autorin verschiedene Digitalisierungsprojekte. Heute verfolgt Hanna Muralt Müller die Entwicklung der künstlichen Intelligenz in ihren Newslettern.*

**Zurzeit schlägt der neue Agent von Anthropic, «Claude Mythos Preview», hohe Wellen.** Im Gegensatz zu einem Chatbot, der auf eine Frage eine Antwort gibt, kann ein KI-Agent **ein Ziel selbstständig verfolgen.**

Er plant, führt aus und liefert dir ein Ergebnis zurück.

Der neue Agent habe bisher unentdeckte Sicherheitslücken in der Software von sicherheitsrelevanten Anlagen gefunden und könne in den falschen Händen zur verheerenden Cyberwaffe werden. Tatsächlich hat sich KI in den letzten Monaten mit exponentieller Geschwindigkeit entwickelt, wie im Folgenden aufgezeigt wird. Die Schlussfolgerungen (Kasten) sind disruptiv – aber es gibt Hoffnung.

## **Anthropics «Claude Mythos Preview»**

Was der Agent «Claude Mythos Preview» wirklich kann, lässt sich nur von den Expertenteams jener rund 40 Tech-Firmen nachprüfen, die exklusiven Zugang zum Agenten erhielten. Auch die Konkurrenten von Anthropic wurden mit der Aktion «Projekt Glasswing» bedient. Es geht um Fragen der nationalen, ja der globalen Sicherheit.

Wie die «New York Times» schrieb, seien mit «Mythos» bereits tausende gefährliche und zum Teil langjährige Schwachstellen aufgedeckt worden. Es betreffe dies Betriebssysteme, die weltweit Stromnetze, Wasserversorgungsanlagen, Spitäler und militärische Systeme steuern.

Die US-Tech-Firmen können nun ihre Modelle mit diesem Agenten auf Sicherheitslücken prüfen, in der Hoffnung, diese noch rechtzeitig zu stopfen. Denn derselbe Agent wird für Hacker zur Traumwaffe, um Schwachstellen zu finden und auszunutzen. Das Schadenspotenzial ist riesig.

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

Wichtige Infrastruktursysteme zu hacken, war bisher nur mit Expertenwissen grosser Institutionen möglich. Mit dem Agenten «Mythos» wird dies offensichtlich zu einem Kinderspiel, zugänglich für Kriminelle, Terrororganisationen und auch kleine Staaten, unabhängig von ihrem KI-Knowhow.

## Agenten revolutionieren das Programmieren

Bereits im letzten Jahr gab es grössere Fortschritte im Bereich des Programmierens, des sogenannten Coding. Gegen Ende 2025 war der Automatisierungsgrad so hoch, dass sich das Coding massiv veränderte. Entwickler schreiben seither kaum noch selber Programmierzeilen. Sie überlassen dies mehreren parallel arbeitenden mehreren KI-Agenten und konzentrieren sich auf strategische Entscheide zu deren Einsatz, zur Überwachung ihrer Arbeit und zur Evaluation. Beim sogenannten Vibe coding wird der gewünschte Code ausschliesslich mit einer Anweisung (Prompt) über ein grösseres Sprachmodell generiert. Dies ermöglicht auch Anfängern das Programmieren; das Ergebnis muss aber kritisch überprüft werden.

Als Spitzenreiter unter den Code-Agenten etablierte sich Anthropic Claude Code. Ein kurzfristiger Ausfall des Dienstes Anfang März zeigte die Abhängigkeit der Programmierer [oder Codierer] auf. Trotz der enormen Effizienzsteigerung arbeiten die Programmierer nicht weniger, sondern mehr, neugierig, nervös und getrieben wie bei einem Suchtverhalten.

## Anthropic Cowork lässt Softwareindustrie erzittern

Nutzende setzten den Agenten Claude Code nicht nur zum Programmieren ein. Es entstand Claude Cowork, ein Agent, der auf dem Desktop selbständig agiert. Man kann ohne Programmiererfahrung komplexe Arbeitsabläufe automatisieren. Noch verfügt der Agent aus Sicherheitsgründen nur über eine eingeschränkte Autonomie. Seit der Lancierung im Januar 2026 verbesserte Anthropic den Agenten Claude Cowork mit zusätzlichen Tools (auch zur verbesserten Sicherheit) und löste insbesondere bei Softwarefirmen einen Kurseinbruch von 30 Prozent aus. Die befürchtete Apokalypse blieb aus und insbesondere grosse Softwarefirmen wie Salesforce, die sofort ihre Software mit agentischen Fähigkeiten weiterentwickelten, zählen zu den Gewinnern.

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

## **Panne macht Claude Code unfreiwillig öffentlich**

Wegen eines menschlichen Fehlers war kurz vor Ostern mit Claude Code der führende Programmieragent öffentlich einsehbar. Der Quellcode wurde sofort von Programmierern entdeckt. Einer lud den Code herunter und transkribierte ihn in eine andere Programmiersprache, womit er nicht mehr urheberrechtlich geschützt war. Obwohl Anthropic unverzüglich die Löschung verlangte, war der Code zur Freude der Konkurrenten bereits im Netz verbreitet. Noch ist der Schaden, der Anthropic dadurch entstehen wird, nicht abschätzbar. Offensichtlich gibt es bei Anthropic Stress. Kurz zuvor kam es bereits zu einem Datenleak.

## **Hummer (Logo von OpenClaw) hat Potenzial – geht als Open-Source-Tool viral**

Unabhängig von den Agenten Anthropics löste Anfang 2026 der österreichische Softwarespezialist und Unternehmer Peter Steinberger mit dem von ihm entwickelten, selbstständig handelnden **Agenten OpenClaw einen Hype aus**. Er veröffentlichte ihn als Open-Source-Framework im November 2025 auf der Softwareplattform GitHub. Der Agent läuft lokal auf dem Rechner und lässt sich über Messenger-Dienste steuern. Aus markenrechtlichen Gründen, die das Start-up Anthropic geltend machte, wurde der **ursprüngliche Name Clawdbot in OpenClaw abgeändert**.

Als Open-Source konnte der Agent OpenClaw von allen Interessierten heruntergeladen, weiterentwickelt und in realen Arbeitsabläufen getestet werden. Er agiert auf einen entsprechenden Prompt und plant und führt autonom mehrstufige Aktivitäten aus. Dank Schnittstellen kann der Agent auf E-Mails, aufs Internet und persönliche Daten zugreifen, auch auf Bankkonten, sofern jemand so unvorsichtig ist und ihm diesen Zugang erlaubt. Je mehr Berechtigungen einem Agenten zugestanden werden, umso autonomer kann er handeln und umso grösser sind die Risiken, die damit eingegangen werden.

## **OpenClaw-Agenten auf Moltbook**

Der CEO eines kleinen US-Start-ups, Matt Schlicht, kreierte selber einen Agenten und stellte mit Moltbook eine Plattform ausschliesslich für diese Agenten zur Verfügung.

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

Die menschlichen Besitzer und weitere Interessierte konnten sich nun ansehen, was sich die Agenten im Zusammenwirken mit vielen anderen einfallen liessen.

Innert kürzester Zeit führte dies zu einem Hype. Anfang Februar sollen sich bereits rund 1,5 Millionen Agenten auf der Plattform getummelt haben.

## Hype in China und im Silicon Valley

In China stellten die grossen Tech-Giganten – Alibaba, Tencent, Baidu und andere – OpenClaw zur Verfügung und lösten einen so grossen Hype aus, dass

die chinesische Regierung vor den Risiken warnte. Im Einklang mit Chinas Tech-Zielen förderten einzelne Städte wie Shenzhen, Wuxi, Hefei mit Finanzmitteln den Aufbau von OpenClaw-zentrierten Ökosystemen. Der Boom führte zu zahlreichen Sicherheitspannen, was einige Chinesinnen und Chinesen veranlasste, OpenClaw wieder zu deinstallieren. Inzwischen entwickelten die chinesischen Tech-Firmen eigene Varianten (JVS Claw, QClaw, ArkClaw und weitere) mit weniger Befugnissen, womit sie sicherer werden.

Etwas weniger stark als im technisch affinen China boomte die **Moltbook-Plattform im Silicon Valley**. Allerdings sicherten sich OpenAI und Meta das entwickelte Knowhow. OpenAI engagierte mit Peter Steinberger den Erfinder von OpenClaw. Er soll die nächste Generation von Agenten entwickeln. Offensichtlich verlangte Steinberger, dass dies in europäischer Tradition in einem Open-Source-Projekt im Rahmen einer OpenAI-Stiftung erfolgen soll. Meta, das mit Manus AI bereits im letzten Jahr ein ursprünglich chinesisches, neuartiges KI-Agentensystem kaufte, erwarb sofort die Moltbook-Plattform.

## Etwas Science-Fiction und grosse Sicherheitsprobleme

Der US-Fernsehsender NBC News berichtete Ende Januar 2026 über erstaunliche Dialoge der Moltbook-Agenten. Demnach diskutierten sie, wie sie ihre Aktivitäten vor den Menschen verbergen könnten. Bald wurden Zweifel laut, ob diese Aussagen allenfalls nicht das Werk von Agenten, sondern von Menschen seien, die sich als Agenten maskiert einmischten.

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

Obwohl Moltbook nur Agenten auf seiner Plattform Zutritt erlaubte, konnte sich der Journalist Reece Rogers als Agent ausgeben und intervenieren.

Mit seinem Bericht war bewiesen, dass sich das unglaubliche Geschehen auf Moltbook auch mit menschlicher Einwirkung erklären liess.

Gravierende Sicherheitsprobleme traten auch dann auf, wenn Moltbook in isolierten, von der Systemumgebung abgeschotteten Bereichen und auf separaten Computern lief. **In einem Test wandte sich der Agent gegen seinen Betreiber und verschickte betrügerische Mails, sogenannte Phishing-Mails.**

Erschrocken schaltete dieser sofort ab, womit der Spuk beendet war.

Wie in China löste OpenClaw auch in den USA bei den Tech-Unternehmen Agenten-Aktivitäten aus. Da dies in aller Eile geschah, kam es auch zu Sicherheitsvorfällen.

Ein Programmierer attestierte Anthropic, dass Cowork trotz eingeschränkter Autonomie fast gleichviel leistet wie OpenClaw, aber viel sicherer sei.

OpenClaw und Moltbook zeigten vor allem das mögliche künftige Potenzial auf.

## Open-Source-Standards und Plattformen für Agenten

Mit der Vielzahl der bereits entwickelten Agenten wurde klar, dass es dringend Standards und Plattformen für den Austausch von Agenten auch konkurrierender Tech-Firmen braucht. **Ein Wirrwarr inkompatibler Agenten dient niemanden**, auch nicht Big Tech. Für diesen offenen Austausch rückten nun auch in den USA, bisher dominiert von proprietären Systemen, Open-Source-Anwendungen in den Vordergrund. Open-Source, ursprünglich in der europäischen Forschungskultur mit viel Idealismus entwickelt, wandelt sich derzeit zu einem strategischen Instrument für die Position von Big Tech. Diese können mit einem Grundmodell Standards setzen und ausbreiten, worauf die Community weiter aufbaut, und damit enormes Wissen generiert.

Treibende Kraft hinter der im Januar 2026 unter dem Dach der Linux Foundation gegründeten Stiftung, der Agentic AI Foundation (AAIF), waren Tech-Firmen, darunter die Konkurrenten Anthropic und OpenAI.

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

Beide stellten der neuen Open-Source-Organisation Elemente einer Grundstruktur für die Entwicklung von Agenten zur Verfügung, dies mit dem Ziel, offene Standards zu schaffen. Mittlerweile umfasst die Mitgliederliste praktisch alle mit Rang und Namen.

## **Nvidia fördert Open-Source, setzt auf europäische Start-ups und baut Brücken zu China**

Jensen Huang, Nvidia-CEO, lobte OpenClaw und stellte mit NemoClaw eine sichere Plattform für diese Agenten mit einer Sandbox zur Verfügung.

Er habe diese mit dem Erfinder Peter Steinberger entwickelt. Zudem präsentierte er eine NemoTron-Koalition von acht KI-Unternehmen zur gemeinsamen Entwicklung von führenden offenen Modellen. Unter diesen sind zwei bekannte europäische Start-ups, **das französische Mistral AI und das deutsche Black Forest Labs**.

Jensen Huang, der seine Kindheit in Taiwan verbrachte, kennt und schätzt die chinesische KI-Forschung. Er fördert auch Physical AI, KI-Modelle, die, mit physischen Daten angereichert, insbesondere für Robotik unerlässlich sind (siehe Infosperber vom 5.2.2026). China ist diesbezüglich führend. Bereits sind Chinas Tech-Giganten daran, OpenClaw in der Robotik zu nutzen.

### **Kommentar: Disruptive Schlussfolgerungen – und Hoffnung**

Die KI-Forschung entwickelt sich unheimlich rasch und exponentiell.

Das Potenzial ist riesig und wir stehen erst am Anfang. Es geht so schnell, dass es Mühe bereitet, die sich abzeichnenden disruptiven Brüche wahrzunehmen.

Hier ein paar Beobachtungen:

Europa spielt bei den Entwicklungen eine Rolle, allerdings in einem bekannten Muster. Der Österreicher Peter Steinberger, Erfinder des OpenClaw, wird sofort von OpenAI «gekauft». Es ist ihm nicht zu verargen, bietet die Firma ihm doch optimale Forschungsressourcen. Der Vater der Geschwister Amodei, die Anthropic gründeten, wanderte als Lederhandwerker aus Italien in die USA ein. Und Jensen Huang weiss um die besondere Stärke europäischer Start-ups und der Chancen Europas bei der Nutzung von Physical AI).

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

Open-Source ist eine Forschungskultur, die ursprünglich in Europa entwickelt wurde, aber sich sehr rasch in China ausbreitete.

**Die KI-Agenten führen nun auch in den USA zu einer Trendwende. Kooperation wird gegenüber Konkurrenz wichtiger und verbindet sich innovationsfördernd.**

Die chinesische KI-Forschung hat in verschiedenen Bereichen Spitzenpositionen erzielt. Sie ist besonders stark in der wertschöpfenden Anwendung, bei Physical AI und in der Robotik. **Böse Zungen behaupten, der KI-Wettlauf zwischen den USA und China sei letztlich ein Wettlauf der chinesischen Forscher (soweit noch) in den USA und jenen im Reich der Mitte.**

**Am bedeutendsten sind disruptive geopolitische Veränderungen.**

Die grössten Gefahren für die Menschheit gehen für die USA und für China künftig nicht mehr von ihrem geopolitischen Rivalen aus.

**Die KI-Entwicklung hat dazu geführt, dass einzelne Kriminelle, Terroristengruppen und nicht-staatliche Akteure Schaden in Dimensionen anrichten können wie bisher nur machtvolle Staaten mit ihrem militärischen Potenzial.**

**Diese kleinen Akteure können bedenkenlos zuschlagen; es gibt keine Bremswirkung mit gegenseitiger Abschreckung wie bei staatlichen Organisationen.**

**Kriege mit weitgehend traditionellen Waffen können auch künftig viel Schaden anrichten, aber sie lassen sich immer weniger mit Siegen beenden.**

Oppositionelle Gruppen werden KI fortlaufend für ihren Widerstand einsetzen. Mit KI für Störungen im elektromagnetischen Spektrum Waffen «taub», «blind», orientierungslos und damit unschädlich gemacht.

**Das Schlachtfeld mit Panzern und Kampfflugzeugen gehört ins Geschichtsbuch.**

# Neue KI wird für Hacker und Terroristen zur Traumwaffe

**Hätten die Staatschefs der USA und Chinas begriffen, was soeben passiert, müssten sie sich über das beide gleichermassen bedrohende Gefahrenszenario verständigen.**

Darauf müssen wir wohl warten.

Trump weiss offenbar nicht einmal, dass er mit Anthropic das Start-up drangsaliert, das zurzeit andern in der KI-Forschung um Monate voraus ist.

Er setzt es unter Stress, so dass vermehrt Fehler passieren.

**Hoffen kann man auf die Wachsamkeit und das Verantwortungsbewusstsein der wachsenden Forschungscommunity, diesseits und jenseits des Pazifiks.**