

Liebe Mitglieder des Netzwerkes der ehemaligen SSAB, liebe Interessierte

Vorfälle mit KI-Agenten und immer leistungsfähigere chinesische Open-Source-Modelle fordern die USA im KI-Wettlauf heraus. Sie müssen erkennen, dass sie bei Open Source im Verzug sind. Zudem übernimmt China auch politisch den Lead mit seiner Initiative für eine globale KI-Governance.

Sich selbstständig organisierende KI-Agenten mit erschreckendem Potenzial

Gemäss einem Bericht der Nachrichtenagentur [Reuters](#) registrierte Taiwan einen völlig neuartigen Hackerangriff. Mit Prompts in Aktion gesetzt, entwickelten KI-Agenten selbstständig ihre Strategie, verbesserten diese fortlaufend, erprobten und evaluierten ihre Angriffspfade in gegenseitiger Kooperation und erzielten Effekte, für die es bisher ganze Teams qualifizierter Hacker brauchte. Es gelang den taiwanischen Behörden, mit den Angriffen zurechtzukommen. Es kündigt sich ein neues Zeitalter der Cyberkriegsführung an.

Sicherheitsvorfälle mit sich selbst steuernden KI-Agenten in den USA

Der Hackerangriff in Taiwan zeigt, wie rasch technische Fortschritte zur ernsthaften Gefahr werden. Nur kurz zuvor wurde in den USA das Potenzial sich selbst organisierender KI-Agenten bei scheinbar harmlosen Tests sichtbar. [KI-Agenten von OpenAI](#) brachen aus der vermeintlich sicheren Testumgebung aus. Sie fanden eine bisher nicht bemerkte Sicherheitslücke und holten sich zur Erfüllung ihres Auftrages Informationen im Internet. Sie starteten hierzu bei [Hugging Face](#), einer der wichtigsten Open-Source-Plattformen, innert kürzester Zeit 17'000 Angriffe. Offensichtlich entdeckte OpenAI den Vorfall erst rund eine Woche später. Es verlor eindeutig die Kontrolle. Und dies war [kein Einzelfall](#). Auch Anthropic und andere KI-Firmen entdeckten anschliessend ähnliche, völlig autonome Aktivitäten, die ausserhalb ihrer Kontrolle abliefen.

Diskussionen in den USA zu Open Source

Bisher gingen die USA davon aus, dass sie mit ihren proprietären Modellen mehr Sicherheit erzielen können als dies bei Open Source möglich sei. Die leistungsfähigsten KI-Modelle von Anthropic und OpenAI – Mythos 5 und GPT-5.6 Sol – wurden vorübergehend vom Netz genommen und stehen aus Sicherheitsgründen zurzeit nur eingeschränkt zur Verfügung. Nun passierten die Sicherheitsvorfälle gerade bei diesen KI-Modellen in den Testphasen und Hugging Face [stoppte](#) den Angriff mit dem heruntergeladenen und selbst gehosteten chinesischen Open-Source-Modell [GLM-5.2](#). Wegen der eingebauten Sicherheitsmassnahmen – welch eine Ironie – verweigerten die US-KI-Modelle einen Gegenangriff zur Verteidigung.

Open Source – die USA richten ihre Strategie neu aus

Ohnehin standen die USA seit Wochen unter dem Schock laufend neuer, sehr leistungsfähiger chinesischer Open-Source-Modelle, die überdies viel kostengünstiger waren als die proprietären US-Modelle. Deshalb sind sie [für die US-Start-ups und die Hochschulforschung unerlässlich](#). China eroberte mit diesen Tools gerade den Weltmarkt. Die künftigen Renditen der proprietären US-Modelle mit ihren Milliardeninvestitionen werden damit bedroht. Auf Vorschläge, die Nutzung dieser chinesischen Start-ups zu beschränken, gab es in den USA eine richtige [Bewegung für Open Source](#). Mit Blick auf den KI-Wettlauf ging es nun darum, den Rückstand auf China aufzuholen.

Jailbreaks sowohl bei Open Source wie bei proprietären Modellen

Gemäss Tests von [FAR.AI](#), einer in Kalifornien basierten gemeinnützigen Organisation, lassen sich sowohl offene wie proprietäre KI-Modelle mit [geschickten Prompts](#) zur Herausgabe gefährlicher Information, zu einem [AI Jailbreak](#), verleiten. Mit den in proprietären Modellen eingebauten Sicherheitsmassnahmen ist die Gefahr reduziert, aber nicht eliminiert. Und weltweit genügt ein einziges Modell, um eine grössere Katastrophe auszulösen.

China positioniert sich mit KI-Governance

Chinas Staatschef Xi Jinping hat begriffen, wie wichtig eine globale KI-Governance ist. Er [gründete](#) im Juli die [World Artificial Intelligence Cooperation Organization](#) (WAICO) mit bereits 29 Mitgliedern vor allem aus dem globalen Süden. China kennt [ethische Richtlinien](#), richtet seine [KI-Regulierung](#) stark auf soziale Stabilität und nationale Sicherheit aus. Es fallen hehre Begriffe, doch mit dem Ziel sozialer Stabilität lässt sich der Überwachungsstaat legitimieren, mit jenem der nationalen Sicherheit können Attacken wie jene auf Taiwan gerechtfertigt werden.

Mit freundlichen Grüssen

Für das Netzwerk der ehemaligen SSAB: Hanna Muralt Müller

24.8.2026

Falls Sie diese E-Mail nicht mehr erhalten möchten, melden Sie sich bitte bei mir: info@muralt-mueller.ch.