

Liebe Mitglieder des Netzwerkes der ehemaligen SSAB, liebe Interessierte

KI droht, die Menschheit auszulöschen: Schlagzeilen mit diesem Inhalt haben heftige Diskussionen in allen Medien ausgelöst. Für die einen handelt es sich um Angstmacherei und um Geschäftsinteressen der KI-Labs, andere sehen reale Gefahren. Dieser Dissens ergibt sich auch daraus, dass vielfach nicht zwischen drei unterschiedlichen Bedrohungen und den zu ergreifenden Massnahmen unterschieden wird.

Erstens: Künftiges Bedrohungspotenzial einer KI ausser Kontrolle

KI-Agenten, die ausser Kontrolle agieren, lösten zum Teil Panik aus. Aus dem [Untersuchungsbericht](#) zum Angriff von KI-Agenten auf die Open-Source-Plattform von Hugging Face (s. Newsletter vom 24.8.2026) ging hervor, wie potent die KI-Agenten sich selbst koordinierten, eigenmächtig weitermachten, obwohl sie feststellten, dass sie ausserhalb ihres Auftrags handelten und sogar Massnahmen trafen, ihre unerlaubten Machenschaften zu verbergen. Das war alarmierend, die Forschenden an der Entwicklungsfront erschranken. Die [Kündigung](#) eines Forschers, der nicht weiter an einer die Menschheit gefährdenden KI-Entwicklung mitwirken wollte, löste Panik aus. Dem Vorschlag von [Dario Amodei](#), das Entwicklungstempo zu verlangsamen, schlossen sich trotz erbitterter Rivalität [Elon Musk](#), [Sam Altman](#) und [Demis Hassabis](#) an – die CEO's der vier fortschrittlichsten KI-Labs waren sich einig. Es war offensichtlich geworden, dass die Sicherheitsmassnahmen der KI-Entwicklung nachhinken und eine KI ausser Kontrolle gefährlich wird.

Der derzeitige Entwicklungsstand gibt keinen Anlass für Panik. Aber beim beschleunigten KI-Entwicklungstempo müsste jetzt, [wie Amodei darlegt](#), die bisher vernachlässigte Sicherheitsforschung sofort intensiviert werden. Aber es gibt Dilemmata. Kein KI-Lab will im US-internen KI-Wettlauf zurückfallen, deshalb müssten alle bereit sein, die Entwicklung so weit zu verlangsamen, bis die Sicherheitsmassnahmen greifen und die KI nicht ausser Kontrolle gerät.

[Trump](#) will wegen des KI-Wettlaufs mit China keine Verlangsamung. Auch China lehnt sie ab, sie sei ein [Manöver](#) zum Abbremsen ihrer Aufholjagd. Immerhin will Trump eine neue [AI-Force mit einem Chef \(AI «Czar»\)](#) für mehr Sicherheit schaffen. Eigentlich gäbe es längst eine besonders wirksame Massnahme für mehr Sicherheit, nämlich die konsequente Durchsetzung der Produkthaftpflicht und hohe Strafen.

Zweitens: Aktuelle Gefahr – KI erleichtert kriminelle Hacking-Angriffe

Die Gefahr liegt hier vor allem bei Open-Source-Modellen. Sobald diese lokal heruntergeladen sind, lassen sie sich dank KI für massenhafte Hacking-Angriffe nutzen. Bei proprietären Modellen können [gefährliche Prompts](#) gestoppt werden und dies ist mehrmals erfolgt, so bei Anfragen zur Entwicklung biologischer Waffen. Hier gibt es weitere Schranken. Kriminelle, die sich trotz der in den meisten KI-Modellen eingebauten Sicherheitsmassnahmen Zugang zur Information verschaffen, brauchen für die Umsetzung im sogenannten «nassen» Labor spezifisches Wissen. Bei diesem zweiten Bedrohungsszenario sind es stets Menschen, die mit krimineller Energie diese Angriffe starten müssen. Derzeit gibt es ein [Wettrennen](#) zwischen Organisationen, die dank KI mehr Sicherheitslücken finden als sie schliessen können, und den Hackern, die mit denselben KI-Systemen immer effizientere Angriffe starten.

Drittens: Gefährliche Missverständnisse und Reaktionen unter Zeitdruck

Wie [CNN](#) kürzlich berichtete, wurde im letzten Moment ein US-Angriff auf ein chinesisches Schiff gestoppt. Das Flugzeug war schon in der Luft, als bemerkt wurde, dass die mit KI erstellte Analyse falsch war, wonach das Schiff Teile eines Atomwaffenprogramms transportiere. Dies hätte fast einen Krieg auslösen können.

Es braucht dringend einen «heissen Draht» zwischen Washington und Beijing, damit nicht falsche Annahmen unter Zeitdruck zur Eskalation führen. Es geht auch um autonome Cyberangriffe oder Fehlfunktionen kritischer Systeme, die missverstanden und zu Kurzschlusshandlungen führen können. Tatsächlich haben der US-Finanzminister Scott Bessent und der chinesische Vizepremier He Lifeng in einem Gespräch im Vorfeld des Treffens von Trump mit Xi Jinping vereinbart, einen solchen [Notfall-Meldemechanismus](#) einzurichten. Die Details sollen an einem nächsten Treffen im November zur Sprache kommen.

Mit freundlichen Grüssen

Für das Netzwerk der ehemaligen SSAB: Hanna Muralt Müller

2.10.2026

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