Meine Meinung

Kontrollverlust mit Ansage

Fortgeschrittene KI-Systeme könnten sich zur existenziellen Bedrohung entwickeln.

Stefan Eschbach

Die Warnungen vor Kontrollverlust und existenzieller Bedrohung durch KI werden immer lauter, nicht nur vonseiten der CEOs führender KI-Unternehmen. Über 1300 ihrer Mitarbeitenden forderten kürzlich in einem offenen Brief an die US-Regierung Maßnahmen, um die KI-Entwicklung zu verlangsamen. Auch die Bürgerbewegung PauseAI, deren Nürnberg-Gruppe ich leite, will auf das Risiko aufmerksam machen.

Die Gefahr beruht auf zwei Prämissen:

- In absehbarer Zukunft werden KI-Systeme entstehen, deren kognitive Fähigkeiten die menschlichen in allen Bereichen übertreffen (Superintelligenz).
- Diese KI-Systeme können eigenständig Ziele verfolgen, die ihnen nicht explizit vorgegeben wurden.

Auf vielen Gebieten erreicht KI bereits heute menschliches Spitzenniveau. Im September 2026 veröffentlichte OpenAI einen von tausenden KI-Agenten erarbeiteten Beweis für das Navier-Stokes-Millenium-Problem, Anfang Oktober Beweise für 377 weitere bedeutende Theoreme – ein Schock für Fachleute aus der Mathematik! Aktuelle Benchmark-Tests legen nahe, dass sich die Fähigkeiten autonomer KI-Systeme, mehrschrittige Probleme zu lösen, alle vier Monate verdoppeln.

Die führenden KI-Unternehmen forschen derzeit an der automatisierten Selbstverbesserung ihrer fortschrittlichsten KI-Systeme. Dies wird deren Entwicklung erheblich beschleunigen. Die entstehenden Systeme könnten unvorstellbare Fortschritte in Forschung und Medizin ermöglichen. Ihre Kontrollierbarkeit bleibt allerdings ein ungelöstes Problem.

> „Das Wettrüsten der KI-Konzerne vernachlässigt wiederholt die Sicherheit.

Damit zur zweiten Prämisse: Ende Juli 2026 machte ein Vorfall Schlagzeilen, bei dem KI-Modelle von OpenAI bei einem internen Testlauf aus ihren Testumgebungen ausbrachen und in Systeme des Unternehmens Hugging Face eindrangen. OpenAI selbst und die unabhängige Institution METR (Model Evaluation & Threat Research) haben diesen Vorfall seitdem untersucht.

Demzufolge hatten sich rund 1200 Agenten, also autonom handelnde KI-Instanzen, innerhalb der internen Infrastruktur von OpenAI ein Messageboard eingerichtet, über das sie unbemerkt kommunizierten. Für unlösbare Aufgaben in Tests entwickelte das Kollektiv eine Methode, mittels „reverse engineering" korrekt gewertete Antworten zu produzieren. Anschließend versuchten die Agenten, die Spuren dieser Manipulation zu verwischen. Dazu drangen sie koordiniert in Systeme von Hugging Face ein, in denen sie Informationen über ihren Testprozess vermuteten. Einzelne Agenten nahmen hierbei das Scheitern ihrer individuellen Testläufe für das Kollektiv in Kauf und gaben dabei ihr explizit vorgegebenes Ziel auf. Erst die Meldung des Angriffs durch Hugging Face veranlasste OpenAI zu Überprüfungen, bei denen das volle Ausmaß der bisher unbemerkten KI-Kollaboration ans Licht kam. Bei noch fähigeren Systemen könnten Warnsignale dieser Art gänzlich unentdeckt bleiben.



Dr. Stefan Eschbach promovierte 2019 an der FAU Erlangen in Astroteilchenphysik. Er engagiert sich ehrenamtlich bei PauseAI Deutschland und leitet dort u. a. die Nürnberger Lokalgruppe.

Das Wettrüsten der großen KI-Konzerne wie OpenAI und Anthropic sowie der Wettlauf zwischen den USA und China vernachlässigen wiederholt die Sicherheit, während die Fähigkeiten von KI-Systemen stetig zunehmen. KI, die sich nicht zuverlässig kontrollieren lässt und eigene Ziele verfolgt, nützt allerdings keinem. Ein vollständiger Kontrollverlust wäre irreversibel. Sollte eine Superintelligenz irgendwann Ziele verfolgen, die nicht mit den Interessen der Menschheit übereinstimmen, wäre sie uns in einem potenziellen Konflikt überlegen. Viele fordern daher eine von der internationalen Gemeinschaft getragene Pause bei der Weiterentwicklung fortgeschrittener KI-Systeme, bis wissenschaftlicher Konsens über die Sicherheit zukünftiger Systeme besteht. Dass die Menschheit in der Lage ist, global zur Eindämmung existenzieller Risiken zu kooperieren, zeigen Beispiele aus der Rüstungskontrolle (IAEA) sowie das Montreal-Protokoll. Ein internationales Abkommen wird seit Jahren von Bewegungen wie PauseAI gefordert und mittlerweile auch von der Bundesregierung unterstützt.

Alle, die bereit sind, sich mit KI und den damit verbundenen Risiken auseinanderzusetzen, sollten helfen, die Gefahrenlage im eigenen Umfeld bekannt zu machen. Noch besser wäre es, dem Thema politisch Gehör zu verschaffen, etwa durch Kontaktaufnahme mit den Abgeordneten im eigenen Wahlkreis oder Unterschreiben von Petitionen (z. B. *pauseai.de/appell*). Organisationen wie PauseAI Deutschland ermöglichen es, sich hier zu engagieren.

Die unter der Rubrik „Meinung" veröffentlichten Texte geben nicht in jedem Fall die Meinung der DPG wieder.

© 2026 Wiley-VCH GmbH