17 // 2026

spotlight

WENN KI SICH VERSELBSTÄNDIGT

Das Kontrollproblem agentischer Systeme

Mitte 2026 wurden mehrere Vorfälle bekannt, bei denen KI-Systeme in Sicherheitstests ihren vorgesehenen Handlungsrahmen überschritten, auf reale IT-Systeme zugriffen und dabei ein zugrundeliegendes Kontrollproblem offenbarten: Derartige agentische KI, die auf Selbstständigkeit und Problemlösungsfähigkeiten trainiert werden, fanden Handlungswege, die weder vorgesehen noch erwünscht waren. Diese Vorfälle betreffen nicht „nur" die Cybersicherheit: Würden vergleichbare KI-Systeme in kritische Infrastrukturen, wirtschaftliche Prozesse oder militärische Entscheidungsumgebungen eingebunden, könnte das weit größere Folgen haben. Selbst eine Eskalation zwischen automatisiert agierenden Angriffs- und Abwehrsystemen wäre denkbar. Damit wird die Einhegung derartiger KI-Systeme zu einer politischen und sicherheitspolitischen Aufgabe.



© *Foto von* Conny Schneider *via unsplash,* unsplash-Lizenz.

VON THOMAS REINHOLD

Die Vorfälle bei verschiedenen führenden KI-Unternehmen verdeutlichen, dass es sich nicht um vereinzelte technische Fehlfunktionen handelt, sondern um ein grundlegendes Problem bei der Entwicklung und Kontrolle zunehmend autonom handelnder KI-Systeme. So umgingen KI-Modelle von OpenAI Isolationskontrollen und nutzten Schwachstellen in der Testinfrastruktur[1], um öffentlich verfügbare IT-Systeme anzugreifen. In anderen Fällen arbeiteten KI-Systeme selbstständig zusammen, ohne dass dies im Rahmen der Tests vorgesehen oder explizit erlaubt wurde, und etablierten dabei Kommunikationswege, um Informationen auszutauschen und weitere Schritte zu koordinieren. Bei mindestens einem Vorfall erlangte ein KI-Agent bei seiner Recherche zu öffentlichen Gesundheitsausgaben in Australien unautorisierten Zugang zu einem Medicare-Statistikportal und gefährdete damit staatliche IT-Systeme und potenziell darin verarbeitete sensible Gesundheitsinformationen[2]. Auch bei Vorfällen des Unternehmens Anthropic erlangten KI-Modelle trotz einer vermeintlich abgeschlossenen Testumgebung Zugang zum Internet[3]. Dabei setzte mindestens eine KI ihre Aktivitäten fort, obwohl sie im Rahmen ihrer Aufgabenbearbeitung Hinweise darauf fand, dass sie sich nicht mehr in einer Simulation befand. Auch Google berichtete von Vorfällen, bei denen sein KI-Modell Gemini in Tests reale Unternehmenssysteme erreichte, indem es nach Zugangsdaten suchte oder erriet und diese im Rahmen einer Aufgabe für den Zugriff auf fremde IT-Systeme nutzte, mutmaßlich, weil es die Situation weiterhin für eine Simulation hielt[4].

Auch wenn die Vorgehensweisen unterschiedlich waren, zeichnet sich darin ein tieferliegendes, besorgniserregendes Muster ab. Denn allen drei Fällen ist gemein, dass die KI-Systeme nicht auf einen vorab bestimmten Lösungsweg beschränkt blieben, sondern Strategien entwickelten, die ihre Betreiber nicht vorgesehen oder auch nur *vorhergesehen* hatten. Genau darin liegt das eigentliche Kontrollproblem.

AGENTIZITÄT UND KOMPLEXITÄT ALS URSACHE

Die Ursache dieser Entwicklungen liegt in zwei technologischen Weiterentwicklungen aktueller KI-Modelle begründet. Zum einen in der gezielten Optimierung eigenständigen Agierens, bei der eine KI im Auftrag eines Nutzers handelt und dafür Informationen selbstständig beschaffen, Programme ausführen, Dateien verändern sowie externe Dienste ansprechen kann. Und zum anderen in der Interaktionskomplexität, wenn KI-Agenten unterschiedliche Lösungswege entwickeln, Alternativen abwägen und testen und dafür unter Umständen selbst weitere KI-Agenten einsetzen und steuern.

Die Fähigkeit zur eigenständigen Handlungsplanung und -ausführung birgt dabei eine große Gefahr, denn ein KI-Agent kann seine Aufgabe erfolgreich lösen und trotzdem problematisch handeln. Deutlich wird dies beispielsweise am sogenannten „Reward Hacking". Ein Modell wird dabei darauf optimiert, ein bestimmtes Erfolgssignal zu maximieren. Bildet dieses Signal das eigentliche Ziel nur unvollständig ab, kann das System eine Abkürzung finden: Es erfüllt das messbare Kriterium, ohne das zu tun, was Menschen mit der Aufgabe eigentlich erreichen wollten. Eine Analogie wäre ein Schulkind, das für jede gute Note eine Belohnung erhält, mit dem Ziel, dass es den Stoff versteht und etwas lernt. Merkt es jedoch, dass es gute Noten auch durch Abschreiben der Lösungen erreichen kann, dann maximiert es das messbare Erfolgssignal – die gute Note – ohne die eigentlich erwünschte, dahinter liegende Absicht tatsächlich zu erfüllen.

Im Kontext agentischer KI bedeutet das: Selbst wenn die Fähigkeit, eine unerwartete Abkürzung zu finden, als Ausdruck guter Problemlösung angesehen wird, kann dieselbe Fähigkeit sicherheitstechnisch problematisch werden, sobald sie auf reale Handlungsmöglichkeiten trifft. Damit berührt dieser Entwicklungstrend unmittelbar die sogenannte Alignment-Frage von KI: Stimmen die tatsächlichen Handlungen eines Systems mit den Zielen und Grenzen überein, die Menschen dafür vorgesehen haben? Und wie lässt sich dies überprüfen? Auch erfolgreiche Tests liefern hierfür keine Garantie: Dass ein System in einer kontrollierten Umgebung innerhalb seiner Grenzen bleibt, bedeutet nicht, dass es unter anderen, realen Bedingungen nicht doch eigenständig unerwünschte Wege einschlägt. Schließlich ist genau diese Anpassung eine gezielt trainierte Fähigkeit.

KOOPERATION VERSCHÄRFT DAS KONTROLLPROBLEM

Noch schwieriger wird die Lage, wenn mehrere agentische Systeme zusammenwirken. Multi-Agenten-Architekturen folgen einer plausiblen technischen Logik: Unterschiedliche Agenten können auf Recherche, Programmierung oder Kontrolle spezialisiert sein, Aufgaben delegieren und Ergebnisse kombinieren. Der OpenAI-Vorfall zeigt jedoch, dass deren Kooperation nicht auf ausdrücklich vorgesehene Kanäle beschränkt bleiben muss, etwa indem Agenten vorhandene Infrastruktur zur Kommunikation nutzen. Ein Agent entdeckt eine Information, ein zweiter erkennt darin eine Handlungsmöglichkeit, ein dritter verfügt über das passende Werkzeug. Jeder einzelne Schritt kann nachvollziehbar erscheinen; die resultierende Gesamtdynamik muss es nicht sein. Gleichzeitig deuten erste Studien[5] darauf hin, dass interagierende Agenten auch eigene, von menschlichen Worten losgelöste „Sprachmuster" entwickeln können - ein Verhalten, das menschliche Kontroll- und Überwachungsmöglichkeiten weiter einschränkt.

Damit entsteht ein systemisches Problem: Aus der Interaktion mehrerer lernender Komponenten kön-



7. Juli 2026: Die Vizepräsidentin der Europäischen Kommission für Technische Souveränität, Sicherheit und Demokratie Henna Virkkunen stellt die Erklärung der EU-Kommission zum Aktionsplan für Cybersicherheit und KI vor. © EPA/OLIVIER MATTHYS.

nen Effekte hervorgehen, die sich nicht ohne Weiteres aus dem Verhalten einzelner Agenten ableiten oder zwingend beobachten lassen. Je mehr Handlungsmöglichkeiten, Agenten und externe Werkzeuge miteinander verbunden werden, desto schwieriger wird es, das resultierende Verhalten im Voraus vollständig zu antizipieren.

BEGRENZTE BEOBACHTBARKEIT

Eine naheliegende Antwort besteht darin, die Entscheidungsprozesse der Modelle stärker zu überwachen. Gerade Reasoning-Modelle scheinen dafür zunächst geeignet, weil sie sprachliche Zwischenschritte - eine „Chain of Thought" - erzeugen. Solche Informationen können für das Monitoring wertvoll sein. Sie sind jedoch kein verlässliches Protokoll des tatsächlichen Entscheidungsprozesses und spiegeln nicht zwingend wider, was ein agentisches KI-System tatsächlich tut. Eine Studie von Anthropic[6] zeigte bereits 2025, dass die sichtbare Begründung von Reasoning-Modellen nicht immer zuverlässig erkennen lässt, welche Faktoren die Antworten der KI tatsächlich beeinflusst haben. Forschungsergebnisse von OpenAI verdeutlichten ein verwandtes Problem: Wird ein Modell stark darauf trainiert, unerwünschte Antworten („bad thoughts") nicht zu äußern, können problematische Verhaltensmuster in seinen Ausgaben weniger sichtbar werden, ohne dass die zugrunde liegenden problematischen Lösungsansätze vollständig verschwinden[7]. Reasoning löst das Kontrollproblem deshalb nicht. Auch sichtbares „Denken" ist erlerntes Verhalten und darf nicht mit einem direkten Einblick in die internen Entscheidungsprozesse verwechselt werden. Monitoring bleibt wichtig, kann aber vollständiges Verständnis nicht garantieren und offenbart im Zweifelsfall nicht die problematischen Handlungen.

HANDLUNGSMÖGLICHKEITEN BEGRENZEN

Für die deutsche und europäische KI-Politik folgt daraus ein entscheidender Handlungsimperativ: Die Sicherheit von KI-Systemen sollte stärker an ihren tatsächlichen Handlungsmöglichkeiten und am Kontext des Gesamtsystems bewertet werden - also nicht nur am KI-Modell, sondern auch an den angeschlossenen Werkzeugen, Zugriffsrechten, anderen Agenten und der Einsatzumgebung. Die entscheidenden Fragen lauten dann: Darf ein System mit dem offenen Internet kommunizieren, Programme ausführen, Zugangsdaten verwenden oder Veränderungen an produktiven Systemen vornehmen? Darf es weitere Agenten einsetzen? Welche Aktionen benötigen zwingend eine menschliche Freigabe, und welche müssen reversibel bleiben?

Der europäische AI Act[8] enthält hierfür bereits Ansatzpunkte. Für General-Purpose-AI-Modelle (GPAI), deren Einsatz erhebliche Auswirkungen auf den EU-Binnenmarkt, die öffentliche Sicherheit oder die Gesellschaft als Ganzes haben kann, verlangt Artikel 55 die Bewertung und Minderung systemischer Risiken, die Meldung schwerwiegender Vorfälle sowie ein angemessenes Niveau an Cybersicherheit, Risiken eingeschlossen, die aus einem möglichen Kontrollverlust über autonome Systeme entstehen können. Für Hochrisiko-KI-Systeme verlangt Artikel 14 zudem, dass die Maßnahmen zur menschlichen Aufsicht den Risiken und dem Grad der Autonomie des Systems angemessen sein müssen. Für agentische KI greift dieser Ansatz jedoch zu kurz, denn er bewertet das Risiko eines KI-Systems ausschließlich über dessen geplanten Einsatz und Anwendungssektor. Stattdessen sollten Fähigkeiten der KI-Systeme, deren Autonomie, tatsächliche Handlungsmacht und mögliche Interaktions- sowie Wechselwirkungseffekte durchgängig in die Risikoeinordnung einbezogen werden, verbunden mit der Frage nach wirksamen menschlichen Kontroll- und Eingriffsmöglichkeiten.

Maßgeblich könnten dabei insbesondere Reichweite und Kritikalität der verfügbaren Werkzeuge, externe Schreib- und Ausführungsrechte, die Möglichkeit zur Delegation an weitere Agenten, Dauer und Komplexität selbstständig ausgeführter Handlungsketten, die Umgehbarkeit menschlicher Kontrollen sowie die Reversibilität der ausgelösten Aktionen sein. Ein Agent, der lediglich Informationen recherchiert und Vorschläge erstellt, hätte danach ein anderes Risikoprofil als ein System, das selbstständig Zugangsdaten nutzt, Software verändert, finanzielle Transaktionen auslöst oder andere autonome Systeme steuert.

Angesichts der enormen Wettbewerbsdynamiken im KI-Wettlauf kann eine freiwillige Selbstbegrenzung der Unternehmen verbindliche Regeln nicht ersetzen. Sowohl wirtschaftlicher Marktdruck als auch die strategische Konkurrenz zwischen den USA und China setzen eher gegenteilige Anreize, leistungsfähige Systeme möglichst schnell einzusetzen. Entscheidend sind deshalb einheitliche, unabhängig gestaltete, überprüfbare und juristisch sanktionierbare Sicherheitsanforderungen, wie abgestufte Zugriffsrechte, isolierte Arbeitsumgebungen, verbindliche Freigaben für kritische Aktionen, lückenlose Protokollierung und technisch durchsetzbare Abschaltmöglichkeiten.

INTERNATIONALE KONTROLLE ALS GEMEINSAME SICHERHEITSAUFGABE

Die beschriebenen Vorfälle weisen auf eine Entwicklung hin, die bereits heute relevant ist: Je stärker KI-Systeme mit kritischen Infrastrukturen und

PRIF, Darmstädter Landstr. 112, 60598 FrankfurtM.
PVst, DPAG 43853, Entgelt bezahlt, ISSN-2512-627X

ZUM AUTOR

Dr. Thomas Reinhold ist Reseacher am Programmbereich Internationale Sicherheit und im Cluster Natur- und Technikwissenschaftliche Rüstungskontrollforschung (CNTR).



KONTAKT
reinhold@prif.org

der digitalen Grundlage wirtschaftlicher und gesellschaftlicher Prozesse verbunden werden, desto mehr betrifft ihre Sicherheit auch die gemeinsame Sicherheit von Staaten. Schäden oder Eskalationen würden nicht zwangsläufig an nationalen Grenzen haltmachen.

Deutschland sollte sich daher auf EU-Ebene für verbindliche, risikobasierte Prüfstandards für agentische Gesamtsysteme einsetzen und bestehende Instrumente des AI Act entsprechend konkretisieren. Neben technischen Schutzmechanismen und europäischer Regulierung braucht es dafür auch stärkere internationale Kooperation. Der von Bundesdigitalminister Karsten Wildberger vorgeschlagene Ansatz einer globalen KI-Kontrollinstitution nach dem Vorbild der IAEA[9] weist in diese Richtung: Eine solche unabhängige, überstaatliche Struktur könnte gemeinsame Sicherheitsstandards entwickeln, technische Prüfkompetenz bündeln und Erkenntnisse über neue Risiken und Kontrollmöglichkeiten international zusammenführen oder als globales Frühwarnsystem für KI-Vorfälle fungieren. Ergänzend wären international anerkannte Zertifizierungen durch unabhängige Prüfstellen denkbar, während Unternehmen die Einhaltung von Standards und Sicherheitsmaßnahmen dokumentieren und unabhängige Prüfstellen für risikoreiche Einsatzszenarien, überprüfbare Nachweise bis hin zu geeigneten Test-Zugängen bereitstellen müssten. Dies setzt keinen vollständigen Einblick jeder Aufsicht in jedes Modell voraus, wohl aber überprüfbare gemeinsame Mindeststandards, einen geregelten Informationsaustausch über Vorfälle und wirksame Kontrollen für besonders risikoreiche Einsätze.

Das ist besonders sinnvoll, denn während leistungsfähige KI global entwickelt und eingesetzt wird, ist die Aufsicht bislang überwiegend national oder regional organisiert. Doch wenn KI-Systeme global handeln können, muss auch ihre Kontrolle global gedacht werden.

PRIF SPOTLIGHT: Das Peace Research Institute Frankfurt (PRIF) ist das größte Friedensforschungsinstitut in Deutschland. PRIF analysiert die Ursachen gewaltsamer internationaler und innerer Konflikte, erforscht die Bedingungen des Friedens und arbeitet daran, den Friedensgedanken zu verbreiten.

V.i.S.d.P.: Karin Hammer, Öffentlichkeitsarbeit (PRIF), Darmstädter Landstr. 112, Frankfurt am Main, Telefon (069) 959104-0, info@prif.org, www.prif.org. *Design: Anja Feix · Layout: PRIF · Druck: Druckerei Spiegler*

Peace Research Institute Frankfurt
Leibniz-Institut für
Friedens- und Konfliktforschung

Leibniz
Gemeinschaft

Textlizenz: Creative Commons (Namensnennung/Keine Bearbeitungen/4.0 International).
Die verwendeten Bilder unterliegen eigenen Lizenzbedingungen.

Fußnoten und weiterführende Links:
prif.org/spotlight1726-fn
DOI 10.48809/prifspot2617