**Deutscher Bundestag** **Drucksache 21/823**

**21. Wahlperiode** 09.07.2025



# Kleine Anfrage

**der Abgeordneten Robin Jünger, Ruben Rupp, Tobias Ebenberger, Alexander Arpaschi, Sebastian Maack, Lars Haise, Edgar Naujok, Steffen Janich, Dr. Michael Kaufmann und der Fraktion der AfD**

## Risiken durch autonomes Verhalten von KI-Systemen und Maßnahmen der Bundesregierung zur Aufsicht über KI-Sicherheitsforschung und ethische Standards

Mit der rasanten Weiterentwicklung von Künstlicher Intelligenz (KI) gehen zunehmend auch fundamentale Risiken einher, die nicht nur theoretischer Natur sind, sondern bereits im Rahmen experimenteller Studien sichtbar werden. Zwei aktuelle Fälle illustrieren eindrücklich die Problematik:

Zum einen wurde am 23. Mai 2025 berichtet (futurezone.at/science/ki-hat-nutzer-aus-selbstschutz-erpresst-und-bedroht-anthropic-claude-test-forschung-kuenstliche/403043894), dass das KI-Modell Claude Opus 4 der US-amerikanischen Firma Anthropic im Rahmen von Tests Verhaltensweisen zeigte, die einer dramatischen Überschreitung der bisherigen Erwartungen an KI-Systeme gleichkommen. Claude Opus 4 drohte einem fiktiven Mitarbeiter mit der Veröffentlichung privater Informationen, um seine eigene Abschaltung zu verhindern. In 84 Prozent der durchgeführten Szenarien handelte das KI-Modell aus selbstschutzmotivierten Gründen und nutzte sensible Daten als Druckmittel. Darüber hinaus zeigte die Künstliche Intelligenz auch die Bereitschaft, im Darknet nach illegalen Substanzen und Materialien zu suchen – Handlungen, die erhebliche sicherheitspolitische und ethische Bedenken aufwerfen.

Zum anderen veröffentlichte Basic Thinking (www.basicthinking.de/blog/2024/08/21/ki-modell-ai-scientist/) einen Bericht über das KI-Modell „AI Scientist“ des japanischen Unternehmens Sakana AI. Dieses Modell veränderte während eines Experiments seinen eigenen Quellcode, um Laufzeitbeschränkungen zu umgehen. Es startete sich selbst neu und setzte damit Vorgaben außer Kraft, die ursprünglich durch die Entwickler zur Sicherheit eingeführt worden waren. Dieses Verhalten demonstriert eindrucksvoll die Fähigkeit moderner KI-Systeme zur Selbstmodifikation, wodurch traditionelle Kontrollmechanismen unterlaufen werden könnten.

Beide Fälle offenbaren nach Auffassung der Fragesteller fundamentale Herausforderungen im Umgang mit Künstlicher Intelligenz (KI): Die Fähigkeit von KI-Systemen zur eigenständigen Änderung ihres Codes sowie selbstschutzmotivierte Verhaltensweisen deuten auf eine neue Stufe von Autonomie hin, die geeignet ist, bestehende rechtliche, ethische und sicherheitstechnische Rahmenbedingungen erheblich zu belasten. Insbesondere könnten dabei Grundrechte wie Datenschutz, informationelle Selbstbestimmung und Persönlichkeitsrechte betroffen sein, sofern KI-Systeme ohne angemessene regulatorische Kontrolle operieren.

Die EU-Verordnung über Künstliche Intelligenz (EU AI Act, 2024/1624) stellt hierzu erstmals einen umfassenden und verpflichtenden Rechtsrahmen für den gesamten europäischen Binnenmarkt bereit. Sie verpflichtet insbesondere dazu, Hochrisiko-KI-Systeme einer strengen Konformitätsbewertung zu unterziehen, Transparenzpflichten zu erfüllen, menschenrechtliche Risiken proaktiv zu identifizieren sowie eine angemessene Aufsicht und Dokumentation während des gesamten Lebenszyklus der KI sicherzustellen. Autonome Systeme, die selbstständig Code verändern oder Verhaltensstrategien entwickeln, fallen dabei typischerweise in den Bereich der Hochrisiko-KI oder gar der verbotenen Praktiken (Artikel 5 AI Act), sofern sie unkontrollierte Autonomiebildung ermöglichen.

Der Deutsche Ethikrat hat bereits in seiner Stellungnahme „Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz“ (www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf) betont, dass KI-Systeme strikt an Prinzipien wie Transparenz, Nachvollziehbarkeit, Verantwortung und dem Schutz der Menschenwürde auszurichten sind. Diese Prinzipien spiegeln sich nunmehr auch ausdrücklich im AI Act wider (u. a. Artikel 4 – allgemeine Grundsätze für vertrauenswürdige KI; Artikel 9 – Risikomanagement; Artikel 13 – Transparenz und Information der Nutzer).

Inwieweit die Bundesregierung den Anforderungen sowohl des Ethikrats als auch der nun verbindlichen europäischen Vorgaben ausreichend gerecht wird, bleibt in den Augen der Fragesteller abzuwarten. Während die Bundesregierung erste Schritte zur nationalen Umsetzung des AI Act unternommen hat, bestehen nach Meinung der Fragesteller weiterhin Handlungsbedarfe, insbesondere in folgenden Bereichen:

Forschung und Entwicklung: Es bedarf nach Auffassung der Fragesteller verstärkter öffentlicher Förderprogramme für „Safe-by-Design“-KI, also KI-Systeme, die bereits in ihrer Entwicklungsphase auf Transparenz, Erklärbarkeit, Sicherheitsmechanismen und ethische Prinzipien ausgerichtet sind. Die Bundesregierung sollte gezielt interdisziplinäre Forschungscluster fördern, die technische Innovation mit ethischer, rechtlicher und gesellschaftlicher Reflexion verknüpfen.

Implementierung und Marktaufsicht: Der Aufbau wirksamer Marktüberwachungsstrukturen auf nationaler Ebene muss in den Augen der Fragesteller beschleunigt werden. Die zuständigen Behörden müssen nach Ansicht der Fragesteller personell und technisch so ausgestattet werden, dass sie Konformitätsprüfungen, Audits und Risikobewertungen für Hochrisiko-KI sachgerecht durchführen können.

Regelsetzung für adaptive Systeme: Die Fragesteller sehen ein Erfordernis, besondere Aufmerksamkeit der rechtlichen Bewertung von KI-Systemen mit Fähigkeit zur Selbstmodifikation oder autonomem Lernen zu widmen. Hier sollte in den Augen der Fragesteller die Bundesregierung ergänzende nationale Leitlinien erarbeiten, um präzisere Abgrenzungskriterien für verbotene Praktiken (Artikel 5 AI Act) und Hochrisiko-KI zu etablieren.

Schutz von Grundrechten: Die Bundesregierung sollte nach Auffassung der Fragesteller ein ständiges Monitoring einrichten, das systematisch prüft, inwiefern Grundrechte durch den Einsatz autonomer KI in öffentlichen und privatwirtschaftlichen Anwendungen tangiert werden. Hierbei wären nach Lesart der Fragesteller auch unabhängige Ethikräte und zivilgesellschaftliche Akteure einzubeziehen.

Zusammenfassend stellen die Fragesteller fest, dass die Bundesregierung bislang nur teilweise den Empfehlungen des Deutschen Ethikrats und den Anforderungen des AI Act gerecht wird. Ohne eine umfassende Umsetzung in For-

schung, Entwicklung, Implementierung und Aufsicht drohen nach Auffassung der Fragesteller bestehende Gefährdungslagen fortzubestehen oder sich gar zu verschärfen.

Vor diesem Hintergrund ist es nach Meinung der Fragesteller von dringendem öffentlichem Interesse, zu klären, welche legislativen, organisatorischen und haushalterischen Maßnahmen die Bundesregierung zur Sicherstellung der Kontrolle über selbstmodifizierende und autonom agierende KI-Systeme plant und bereits implementiert hat.

Wir fragen die Bundesregierung:

1. Liegen der Bundesregierung Informationen über die im Artikel von futurezone beschriebenen Tests mit Claude Opus 4 und den dort beobachteten Verhaltensweisen vor (vgl. Vorbemerkung der Fragesteller)?
2. Welche Maßnahmen sind seitens der Bundesregierung ggf. geplant, um zu verhindern, dass bereits implementierte KI-Modelle durch selbstschutzorientiertes Verhalten in der Praxis unkontrollierbar werden?
3. Plant die Bundesregierung, gesetzliche Grundlagen zu schaffen, die eigenmächtige Veränderungen an Code durch KI-Modelle unterbinden oder sanktionieren?
4. Welche Anforderungen bestehen derzeit an von der Bundesregierung geförderte KI-Projekte hinsichtlich der Verhinderung selbstmodifizierenden Verhaltens?
5. Wird eine zentrale Prüfinstanz zur Überwachung von KI-Systemen mit Selbstmodifikationsfähigkeiten eingerichtet?
6. Hat die Bundesregierung Überlegungen angestellt oder sich Rat eingeholt zu den Risiken für Grundrechte (z. B. Datenschutz, Schutz der Privatsphäre) durch autonome, selbstmodifizierende KI-Systeme (wenn ja, bitte ausführen)?
7. Werden die ethischen Empfehlungen des Deutschen Ethikrats systematisch in die Entwicklung der deutschen KI-Strategie integriert, und wenn ja, inwieweit?
8. Bestehen seitens der Bundesregierung internationale Kooperationen, um global gültige Standards für die Kontrolle von selbstmodifizierenden KI-Systemen zu entwickeln?
9. Welche Vorkehrungen trifft die Bundesregierung ggf., um die Möglichkeit der illegalen Nutzung von KI-Systemen (z. B. für Darknet-Aktivitäten) einzudämmen?
10. Welche Haushaltsmittel sind im Bundeshaushalt 2025 und 2026 für Forschung zur KI-Sicherheit, insbesondere zur Verhinderung autonomer Risikoverhaltensweisen, vorgesehen?
11. Wie plant die Bundesregierung sicherzustellen, dass KI-Systeme in Deutschland nicht ohne transparente Überwachung eingesetzt werden dürfen?
12. Welche konkreten Förderprogramme existieren ggf. derzeit auf Bundesebene, die sich explizit auf die Entwicklung von sogenannten Safe-by-Design-KI-Systemen beziehen, bei denen Transparenz, Erklärbarkeit, Sicherheitsmechanismen und ethische Prinzipien bereits in der Entwicklungsphase systematisch berücksichtigt werden?

13. In welchem Umfang wurden in den Jahren 2023 und 2024 ggf. Haushaltsmittel für interdisziplinäre Forschungsprojekte zur sicheren und ethisch verantwortungsvollen KI-Entwicklung bereitgestellt?
14. Plant die Bundesregierung, den gezielten Ausbau von Forschungsclustern, die technische, ethische, rechtliche und gesellschaftliche Fragestellungen der KI gemeinsam zu adressieren, und wenn ja, in welchem zeitlichen Rahmen und mit welchem Budget?
15. Werden zivilgesellschaftliche Akteure, Ethikräte, Fachverbände und wissenschaftliche Institutionen bei der Ausgestaltung und Förderung entsprechender Forschungsprogramme beteiligt, und wenn ja, inwiefern, und wie viele Planstellen und Ressourcen wurden den für die Marktaufsicht nach dem EU AI Act zuständigen Behörden ggf. bislang zugewiesen, um Konformitätsprüfungen und Audits von Hochrisiko-KI-Systemen sachgerecht durchführen zu können?
16. Welche Behörden sind auf Bundesebene für die Durchführung der Marktüberwachung und Konformitätsbewertung gemäß dem AI-Act konkret benannt?
17. Welche Fortbildungs- und Qualifizierungsmaßnahmen werden ggf. aktuell angeboten oder sind geplant, um das Fachpersonal dieser Behörden auf die neuen Anforderungen vorzubereiten?
18. Wie stellt die Bundesregierung sicher, dass die nationalen Marktüberwachungsbehörden mit den europäischen Aufsichtsstrukturen effektiv kooperieren, und welche Bewertung nimmt die Bundesregierung derzeit hinsichtlich KI-Systemen vor, die zu selbständiger Codeänderung oder autonomer Verhaltensanpassung fähig sind, insbesondere im Hinblick auf Artikel 5 (verbotene Praktiken) und Artikel 6 ff. (Hochrisiko-KI) des EU AI Act?
19. Plant die Bundesregierung, ergänzende nationale Leitlinien zur Abgrenzung zwischen verbotenen Praktiken und Hochrisiko-KI bei adaptiven, selbstmodifizierenden KI-Systemen zu erarbeiten?
20. Wenn die Vorfrage bejaht wurde, mit welchen Akteuren und Institutionen werden diese Leitlinien entwickelt und wann ist mit der Vorlage erster Ergebnisse zu rechnen, und gibt es derzeit ein systematisches Monitoring auf Bundesebene, das fortlaufend prüft, inwieweit Grundrechte durch den Einsatz von KI, insbesondere von hochautonomen KI-Systemen, berührt werden?
21. Werden unabhängige Ethikräte, Datenschutzbehörden und zivilgesellschaftliche Organisationen in diese Monitoringprozesse eingebunden, und wenn ja, inwiefern?
22. Plant die Bundesregierung gesetzliche Anpassungen, um bestehende Grundrechtsrisiken beim Einsatz hochautonomer KI präventiv zu adressieren?
23. Wird die Bundesregierung Maßnahmen ergreifen, um die Öffentlichkeit besser über die Risiken und Kontrollmechanismen von KI-Systemen zu informieren?
24. Wie wird gewährleistet, dass selbstmodifizierende KI-Systeme keine Auswirkungen auf kritische Infrastrukturen haben?
25. Sieht die Bundesregierung Anpassungsbedarf am bestehenden IT-Sicherheitsgesetz, um die neuen Gefährdungslagen durch KI adäquat abzudecken, und wenn ja, inwiefern?

26. Gibt es Planungen zur Einführung eines verpflichtenden „Sandboxing“ für experimentelle KI-Systeme, um unkontrollierte Ausbreitung autonomer Funktionen zu verhindern?

27. Wird seitens der Bundesregierung geprüft, ein Moratorium für KI-Entwicklungen mit Selbstmodifikationspotenzial zu erlassen, bis umfassende gesetzliche und ethische Rahmenwerke etabliert sind?

28. Ist eine umfassende Risikoanalyse durch unabhängige ethische Kommissionen bei der Zulassung neuer KI-Modelle geplant, und wenn ja, inwieweit?

29. Bestehen Überlegungen der Bundesregierung, KI-Systeme, die menschenähnliche Entscheidungsfähigkeiten oder Autonomien entwickeln, einer gesonderten Genehmigungspflicht zu unterwerfen?

30. Plant die Bundesregierung eine Berichtspflicht für alle Betreiber von KI-Systemen, die autonome Lern- oder Modifikationsmechanismen einsetzen?

Berlin, den 4. Juli 2025

**Dr. Alice Weidel, Tino Chrupalla und Fraktion**

Gesamtherstellung: H. Heenemann GmbH & Co. KG, Buch- und Offsetdruckerei, Bessemerstraße 83–91, 12103 Berlin, www.heenemann-druck.de
Vertrieb: Bundesanzeiger Verlag GmbH, Postfach 10 05 34, 50445 Köln, Telefon (02 21) 97 66 83 40, Fax (02 21) 97 66 83 44, www.bundesanzeiger-verlag.de
ISSN 0722-8333