

Antrag

der Abgeordneten Rebecca Lenhard, Dr. Konstantin von Notz, Jeanne Dillschneider, Dr. Anna Lührmann, Dr. Moritz Heuberger, Karl Bär, Dr. Franziska Brantner, Dr. Irene Mihalic, Lukas Benner, Sara Nanni, Marlene Schönberger, Dr. Till Steffen, Filiz Polat, Dr. Andrea Lübcke und der Fraktion BÜNDNIS 90/DIE GRÜNEN

KI-Sicherheit stärken – Risiken leistungsfähiger KI-Systeme wirksam begrenzen

Der Bundestag wolle beschließen:

I. Der Deutsche Bundestag stellt fest:

Die Entwicklung Künstlicher Intelligenz (KI) hat allein im letzten Jahr enorme Fortschritte gemacht. KI-Systeme werden nicht nur immer leistungsfähiger, sie übernehmen zunehmend komplexe Aufgaben eigenständig. Fortschritte bei autonomen KI-Systemen verschieben dabei aber auch den Umfang möglicher Risiken immer weiter. Die öffentliche Debatte über KI-Sicherheit wird derzeit stark von Aussagen der großen KI-Unternehmen und ihrer Führungskräfte geprägt. Dadurch richtet sich der Blick schnell vor allem darauf, wie leistungsfähig oder „mächtig“ ihre Systeme werden könnten, während konkrete Versäumnisse bei Sicherheitsvorkehrungen und die Verantwortung der Unternehmen selbst in den Hintergrund geraten. Für politische Entscheidungen müssen deshalb unabhängige wissenschaftliche Erkenntnisse und Einschätzungen von Sicherheitsforscherinnen und Sicherheitsforschern ebenso maßgeblich sein wie die konkreten Vorfälle, die bereits heute zeigen, wo Schutzmaßnahmen nicht ausreichen.

Bei Sicherheitstests führender KI-Labore sind KI-Agenten wiederholt vorgesehene Begrenzungen umgangen, haben unerlaubt auf das Internet zugegriffen, reale Systeme Dritter erreicht sowie auf diese unrechtmäßig Zugang erlangt (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>). Besonders bedenklich ist dabei die Entwicklung von Systemen, die sich selbst oder andere KI-Systeme zunehmend eigenständig trainieren und verbessern. Wenn KI-Systeme durch zunehmende Selbstständigkeit menschliche Aufsicht gezielt umgehen oder sich einer Abschaltung entziehen können, steigt das Risiko eines Verlusts menschlicher Kontrolle mit heute noch weitgehend unabsehbaren Folgen enorm. Solche Risiken betreffen zunehmend auch die staatliche Verwaltung. Durch den fortschreitenden Einsatz von KI-Systemen, aktuell beispielsweise für den Migrationsbereich konkret vorgesehen (BT-Drs. 21/8275), entstehen neue Gefahrenpotenziale im Hinblick auf Verwaltungsentscheidungen, die effektive und grundrechtsschonende Sicherheitsmechanismen dringend erforderlich machen.

Neben den Risiken steigen auch Umfang und Schwere möglicher Schäden. KI-Systeme werden schon heute für Cyberangriffe, Manipulation und Betrug missbraucht. Expertinnen und Experten sehen zudem erhebliche Risiken beim Einsatz solcher Systeme in biologischen, chemischen und anderen besonders sicherheitskritischen Bereichen. Das sind keine abstrakten Probleme, sondern reale Gefahren für die öffentliche Sicherheit und kritische Infrastrukturen, wie es auch EU-Kommissionspräsidentin Ursula von der Leyen in ihrer Rede zur Lage der Union deutlich benannt hat (https://ec.europa.eu/commission/presscorner/detail/de/SPEECH_26_1868).

Cyberangriffe verursachen bereits heute enorme wirtschaftliche Schäden: Allein durch weitreichende IT-Angriffe, Spionage und Sabotage entstehen jährlich Schäden in dreistelliger Milliardenhöhe, wobei Cyberangriffe für einen Großteil dieser Schäden verantwortlich sind (<https://www.bitkom.org/sites/main/files/2026-09/bitkom-studienbericht-wirtschaftsschutz-2026.pdf>). Mit dem zunehmenden Einsatz von KI können Angriffe schneller, skalierbarer und zielgerichteter werden. Die Begrenzung KI-bedingter Sicherheitsrisiken dient deshalb auch dem Schutz der deutschen Wirtschaft und ihrer Wettbewerbsfähigkeit.

Die antragstellende Fraktion hat bereits frühzeitig auf neue Sicherheitsrisiken durch die zunehmenden Fähigkeiten von KI-Systemen hingewiesen. In ihrem Antrag „Vertrauenswürdige Künstliche Intelligenz ermöglichen“ (BT-Drs. 21/2349) hat sie insbesondere auf neue Risiken durch KI-Agenten für die IT-Sicherheit aufmerksam gemacht und eine Überprüfung des bestehenden Rechtsrahmens gefordert. Mit dem Antrag „Europa zum KI-Standort machen“ (BT-Drs. 21/6825) hat sie zudem aufgezeigt, dass sichere und vertrauenswürdige KI zugleich zu einem wichtigen europäischen Qualitätsmerkmal und echtem Standortvorteil werden kann. Konkrete Maßnahmen für einen wirksamen, einheitlichen Schutz kritischer Infrastrukturen hat sie im Antrag „Deutschland resilient machen“ (BT-Drs. 21/2725) vorgelegt.

Zunehmend setzt sich die Erkenntnis durch, dass es ein schwerer Fehler war, dass Politik und bestehende Aufsichtsstrukturen das Heft des Handelns sehr weitgehend den Unternehmen überlassen haben, deren Führungen vielfach dokumentiert haben, dass sie – bedingt durch einen stark ökonomisch geprägten Wettlauf – gesellschaftspolitisch unverantwortlich handeln. Die bereits beschriebenen aufgetretenen Probleme sind gerade darauf zurückzuführen, dass nicht ausreichend Zeit und Kapazitäten in die Absicherung der Systeme fließen. Zudem werden Vorfälle nur partiell, selektiv und verspätet Betroffenen wie auch zuständigen Behörden gemeldet.

Es ist überfällig, das Heft des Handelns zurückzuerobern, bestehende Kontroll- und Sicherungsregime zu effektivieren und die sie durchsetzenden Aufsichtsbehörden entsprechend zu stärken. Deutschland und Europa müssen jetzt Verantwortung übernehmen und dafür sorgen, dass aktuelle Risiken frühzeitig erkannt, unabhängig bewertet und wirksam begrenzt werden. Dazu müssen bereits bestehende Risiken heutiger Systeme wirksamer begrenzt und zugleich Sicherheitsvorkehrungen für Fähigkeiten getroffen werden, die sich derzeit rasant entwickeln. Die Sicherheit von KI-Anwendungen darf nicht länger allein den Unternehmen überlassen werden, die diese Systeme entwickeln. Wenige marktmächtige Anbieter dürfen nicht faktisch selbst darüber entscheiden, welche Risiken gesellschaftlich akzeptabel sind und welche Sicherheitsstandards zukünftig gelten. Es kann nach Meinung der antragstellenden Fraktion nicht Anspruch des Gesetzgebers sein und sollte auch gesellschaftlich nicht akzeptiert

werden, darauf degradiert zu werden, alle paar Monate in neuen Firmenberichten nachlesen zu müssen, wie knapp man erneut an einer Katastrophe vorbeigeschrammt ist.

Es braucht daher unabhängige Prüfungen, verbindliche Sicherheitsanforderungen und staatliche sowie wissenschaftliche Institutionen, die technisch und personell in der Lage sind, mit der weiteren Entwicklung von KI-Anwendungen tatsächlich Schritt zu halten. Bei besonders hohen und nicht hinreichend beherrschten Risiken muss es auch möglich sein, die Veröffentlichung und Bereitstellung eines Modells zu beschränken und die Weiterentwicklung so lange auszusetzen, bis wirksame Schutzmaßnahmen geschaffen wurden. Sicherheitsstandards müssen auf wissenschaftlicher Evidenz beruhen, demokratisch legitimiert, transparent und unabhängig überprüfbar sein. Angesichts der Geschwindigkeit der Entwicklung und des möglichen Schadensausmaßes braucht es jetzt eine erhebliche politische und institutionelle Kraftanstrengung für mehr KI-Sicherheit.

Die Europäische Union und ihre Mitgliedstaaten haben sich mit dem AI Act und dessen Implementierung auf nationalstaatlicher Ebene bei der notwendigen rechtsstaatlichen Regulierung von KI auf den Weg gemacht. Dieser Weg der Regulierung muss konsequent fortgesetzt werden. Nationale Alleingänge oder ein allein europäisches Vorgehen reichen jedoch für diese Risikolage nicht aus. Vielmehr muss die Erhöhung von KI-Sicherheit als internationale Gemeinschaftsaufgabe verstanden und mit einer gemeinsamen globalen Kraftanstrengung vorangebracht werden. Gerade vor dem Hintergrund, dass die derzeit führenden KI-Anwendungen maßgeblich von amerikanischen und chinesischen Unternehmen bereitgestellt werden, wird es für die Etablierung wirksamer internationaler Schutzmechanismen entscheidend sein, dass sowohl die USA als auch China diese mittragen. Dazu braucht es intensive diplomatische Bemühungen und – auch mit Blick auf bisher stark divergierende Regulierungsansätze - vermutlich einen langen Atem. Parallel muss die EU zwingend weitere, konkrete Schritte unternehmen, um (immer neue) Risiken zu minimieren, solange es keine global tragfähige Regulierung gibt.

Bei der Regulierung von KI beweist sich, was Sonntagsreden zur notwendigen Erhöhung staatlicher Souveränität im Digitalen wert sind. Anspruch von Politik muss sein, staatlicher Schutzverantwortung auch auf diesem Gebiet gerecht zu werden und eine sichere, nachhaltige und gemeinwohlorientierte Digitalpolitik umzusetzen.

II. Der Deutsche Bundestag fordert die Bundesregierung auf,

1. ein umfassendes Maßnahmenpaket zur Stärkung der Resilienz von Staat, Wirtschaft und Gesellschaft gegenüber neuen KI-bedingten Sicherheitsrisiken auf den Weg zu bringen und
 - a) insbesondere die personellen, finanziellen und technischen Kapazitäten des Bundesamtes für Sicherheit in der Informationstechnik (BSI) kurzfristig und bedarfsgerecht auszubauen, damit dieses den durch leistungsfähigere und zunehmend autonome KI-Systeme entstehenden neuen Sicherheitsrisiken wirksam begegnen kann;
 - b) das deutsche KI-Sicherheitsinstitut AISI zeitnah zu einer leistungsfähigen und unabhängigen Evaluierungseinrichtung auszubauen, dauerhaft

- ausreichend Personal, finanzielle Mittel und technische Infrastruktur bereitzustellen und eine enge Kooperation des AISI mit dem EU AI Office sowie weiteren europäischen und internationalen KI-Sicherheitsinstitutionen zu gewährleisten;
- c) die Anwendung und Durchsetzung des Cyber Resilience Act (CRA) konsequent sicherzustellen, um die Sicherheit entlang der gesamten Lieferkette zu stärken;
 - d) öffentliche Verwaltung, Betreiber kritischer Infrastrukturen und deutsche Unternehmen, insbesondere kleine und mittlere Unternehmen (KMU), umfassend auf neue KI-bedingte Sicherheitsrisiken vorzubereiten, hierfür gemeinsam mit den Ländern Beratungsstrukturen zu fördern, praxisnahe Handlungsempfehlungen und Notfallkonzepte zu entwickeln und den Aufbau entsprechender Kompetenzen und Krisenreaktionsstrukturen zu unterstützen;
 - e) sicherzustellen, dass unabhängige wissenschaftliche Expertise systematisch und dauerhaft in die staatliche Aufsicht und Sicherheitsbewertung von KI-Systemen einbezogen wird;
 - f) die Bevölkerung umfassend und verständlich über neue, durch KI verschärfte Cyberrisiken und konkrete Schutzmöglichkeiten zu informieren und bestehende Informations-, Beratungs- und Unterstützungsangebote entsprechend auszubauen;
2. sich auf europäischer Ebene dafür einzusetzen, dass die bestehenden Sicherheitsvorgaben für KI-Modelle mit allgemeinem Verwendungszweck mit systemischem Risiko konsequent durchgesetzt und die bestehenden Untersuchungs-, Evaluierungs- und Sanktionsbefugnisse des EU AI Office bei möglichen Verstößen tatsächlich genutzt werden;
 3. zu prüfen, ob die neben der KI-Verordnung geltenden sektorspezifischen Regelwerke neu auftretende und sich verändernde Risiken durch den Einsatz von KI in den jeweiligen Bereichen angemessen berücksichtigen und bestehende Schutz- und Sicherheitsanforderungen sowie Meldepflichten bei Sicherheitsvorfällen bei Bedarf entsprechend weiterzuentwickeln;
 4. verbindliche Sicherheitsanforderungen für autonome und sich selbst oder andere KI-Systeme weiterentwickelnde Systeme voranzutreiben und dabei insbesondere zu prüfen, wie als Ultima Ratio wirksame technische Möglichkeiten zur Unterbrechung und sicheren Abschaltung solcher Systeme gewährleistet werden können;
 5. eine verbindliche und transparente Aufarbeitung schwerwiegender KI-Sicherheitsvorfälle sicherzustellen und die Anbieter der Systeme und deren Führungskräfte durch neue und bereits existierende und praxistaugliche Haftungsregelungen klar zur Verantwortung zu ziehen;
 6. sich auf europäischer und internationaler Ebene, insbesondere im Rahmen der Vereinten Nationen (UN), für gemeinsame KI-

Sicherheitsstandards und eine unabhängige internationale Aufsicht einzusetzen.

Berlin, den 13. Oktober 2026

Katharina Dröge, Britta Haßelmann und Fraktion

Begründung

Die zunehmenden Fähigkeiten und die wachsende Autonomie von KI-Systemen schaffen neue Sicherheitsrisiken und erfordern entschlossenes politisches Handeln. Es ist daher zu begrüßen, dass KI-Sicherheit zunehmend öffentliche und politische Aufmerksamkeit erhält und der Handlungsbedarf stärker in den Fokus rückt. Dabei haben Denkschulen wie Effective Altruism und Longtermism erheblichen Einfluss auf Teile des Diskurses über KI-Sicherheit. (<https://pmc.ncbi.nlm.nih.gov/articles/PMC11912955>). Die Konzentration auf langfristige, hypothetische und besonders dramatische Zukunftsszenarien darf jedoch nicht dazu führen, bereits heute bestehende Schäden, ungleiche Machtverhältnisse und die Verantwortung der Unternehmen aus dem Blick zu verlieren. Eine verantwortungsvolle KI-Sicherheitspolitik muss deshalb sowohl künftige Risiken ernst nehmen als auch die Grundrechte, Sicherheitsbedürfnisse und konkreten Auswirkungen auf Menschen von heute berücksichtigen. Entscheidend ist daher, konkrete Entwicklungen und Vorfälle genau zu betrachten und politische Entscheidungen auf unabhängige wissenschaftliche Erkenntnisse zu stützen.

Zunehmend häufen sich Vorfälle, bei denen KI-Agenten führender KI-Unternehmen vorgesehene Begrenzungen umgangen, unerlaubt auf das Internet zugegriffen, reale Systeme Dritter erreicht sowie auf diesen unrechtmäßig Zugang erlangten (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>). Immer wieder werden solche, durchaus schwerwiegende Vorfälle erst im Nachhinein entdeckt. Auch nachträgliche Analysen können nicht mit vollständiger Sicherheit ausschließen, dass weitere durch KI-Agenten vorgenommene Schritte bislang unentdeckt geblieben sind. Die vielfältigen Vorfälle, die allein in jüngster Zeit öffentlich wurden, zeigen deutlich, wie schwierig es schon heute ist, Risiken beim Einsatz autonomer KI-Systeme zu minimieren, sie vollständig zu überwachen und ihre Handlungen zuverlässig einzugrenzen. Hinzu kommt, dass sich Leistungsfähigkeit und Autonomie solcher Systeme derzeit mit einer Dynamik entwickeln, die teilweise selbst frühere Erwartungen von Fachleuten übertrifft (<https://metr.org/time-horizons/>).

Immer wieder kam es auch zu Fehlern auf Seiten der KI-Unternehmen, die den Einsatz ihrer Anwendungen durch Dritte nicht ausreichend kontrollierten, was unter anderem dazu führte, dass diese selbst bei kriegesischen Auseinandersetzungen zum Einsatz kamen (<https://www.anthropic.com/threat-intelligence-report-september-2026#conventional-weapons-sep-26>). Im Extremfall können Fehlentwicklungen oder der gezielte Missbrauch besonders leistungsfähiger KI-Systeme erhebliche Gefahren für die öffentliche Sicherheit und kritische Infrastrukturen verursachen. Auch an der Schnittstelle von KI und Biotechnologie können neue Sicherheitsrisiken entstehen. Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) hat hierzu 2026 mögliche Risiken für die Biosicherheit und die Proliferation biologischer Waffen untersucht. Die Ergebnisse wurden am 8. Juli 2026 in einem öffentlichen Fachgespräch des Deutschen Bundestages mit Expertinnen und Experten diskutiert (<https://www.tab-beim-bundestag.de/news-2026-08-ki-und-biotechnologie-neue-tab-studie-analysiert-sicherheitsrisiken-und-handlungsoptionen.php>).

Eine wichtige Rolle bei der Aufdeckung von Sicherheitsproblemen bei KI-Systemen spielten bei den letzten Fällen auch aktuelle und ehemalige Beschäftigte der KI-Unternehmen, die aufgrund ihrer Tätigkeit über besondere Einblicke in die Entwicklung der Systeme und unternehmensinterne Sicherheitsvorkehrungen

Vorabfassung – wird durch die lektorierte Fassung ersetzt.

verfügen. Umso wichtiger ist ein wirksamer Hinweisgeberschutz, der es Beschäftigten ermöglicht, mögliche Verstöße gegen bestehende Sicherheitsvorgaben gegenüber den zuständigen Stellen zu melden, ohne Repressalien befürchten zu müssen.

Besonders relevant ist die neue Fähigkeit von KI-Systemen, sich eigenständig zu verbessern, sogenanntes recursive self-improvement (RSI). Entsprechende Warnungen vor gravierenden Risiken kommen dabei nicht nur von den medienwirksam auftretenden Führungskräften der großen Technologieunternehmen, sondern ebenso aus der unabhängigen Forschung und von Vertreterinnen und Vertretern relevanter Kontrollinstanzen (<https://www.aisi.gov.uk/frontier-ai-trends-report>). Auch vor diesem Hintergrund können diese Warnungen nicht allein als wirtschaftlich oder strategisch motivierte Kommunikation der Unternehmen abgetan werden. Was jedoch zutrifft: Die weitere Dynamik ist heute auch insofern nur schwer konkret vorherzusagen, als dass Risikobewertungen und Meldungen über Vorfälle bislang weitgehend der Kommunikation der Unternehmen überlassen wird.

Zu 1. Die zunehmenden Fähigkeiten von KI-Systemen verändern auch die Anforderungen an die staatliche Sicherheitsarchitektur. Deutschland muss deshalb seine bestehenden Strukturen kurzfristig auf neue KI-bedingte Sicherheitsrisiken vorbereiten. Die nationalen Aufsichtsbehörden wie die Bundesnetzagentur (BNetzA), der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit und das Bundesamt für Sicherheit in der Informationstechnik (BSI) müssen über die notwendigen personellen, finanziellen und technischen Kapazitäten verfügen, um neuen Cyberrisiken wirksam begegnen zu können. Zugleich braucht Deutschland mit dem AISI eine leistungsfähige Einrichtung, die die Sicherheit besonders leistungsfähiger KI-Systeme unabhängig untersuchen und bewerten kann. Dafür sollte das AISI eng mit weiteren Forschungseinrichtungen wie dem Institut für KI-Sicherheit des DLR sowie mit deutschen Sicherheitsbehörden und der Bundeswehr zusammenarbeiten. Staat, Wirtschaft und Betreiber kritischer Infrastrukturen müssen sich auf mögliche neue Angriffsszenarien vorbereiten und über geeignete Krisenreaktionsstrukturen verfügen. Ergänzend müssen unabhängige Forschungsstrukturen langfristig gestärkt werden, um Sicherheit, Kontrollierbarkeit und Evaluierung zunehmend autonomer KI-Systeme wissenschaftlich weiterzuentwickeln. Auch Bürgerinnen und Bürger benötigen verständliche Informationen über neue Risiken und konkrete Schutzmöglichkeiten. Zum ganzheitlichen Schutz kritischer Infrastrukturen hat die antragstellende Fraktion bereits im November 2025 einen umfassenden Antrag vorgelegt (BT-Drs. 21/2725).

Zu 2. Mit der europäischen KI-Verordnung (AI Act) besteht bereits ein verbindlicher Rechtsrahmen für KI-Modelle mit allgemeinem Verwendungszweck mit systemischem Risiko. Das EU AI Office verfügt über Befugnisse zur Untersuchung möglicher Verstöße und zur Evaluierung entsprechender Modelle. Entscheidend ist deshalb, dass diese bestehenden Vorgaben und Befugnisse bei konkreten Verdachtsfällen konsequent überprüft und tatsächlich angewandt werden.

Zu 3. Neben der KI-Verordnung gelten für den Einsatz von KI zahlreiche sektorspezifische Sicherheitsanforderungen. Angesichts der schnellen Entwicklung der Fähigkeiten und zunehmenden Autonomie von KI-Systemen muss fortlaufend geprüft werden, ob diese Regelwerke neu auftretende Risiken ausreichend berücksichtigen oder bestehende Schutzanforderungen weiterentwickelt werden müssen.

Zu 4. Systeme, die eigenständig handeln, auf externe Systeme zugreifen oder zur Entwicklung, zum Training oder zur Verbesserung weiterer KI-Systeme beitragen können, erfordern besondere Schutzvorkehrungen. Dazu gehören insbesondere abgestufte Zugriffsrechte nach dem Prinzip der geringstmöglichen Berechtigung, abgesicherte Ausführungsumgebungen, vollständige Protokollierung, kontinuierliche Überwachung, technische Möglichkeiten zum Abbruch laufender Prozesse und menschliche Freigaben bei besonders folgenreichen Handlungen. Auch international werden entsprechende Notfallmechanismen zunehmend diskutiert. Der US-Bundesstaat Kalifornien prüft derzeit regulatorische Vorgaben für technische Möglichkeiten zur Notabschaltung besonders leistungsfähiger KI-Modelle sowie für eine unabhängige Überprüfung ihrer Wirksamkeit (<https://www.gov.ca.gov/2026/09/18/governor-newsom-issues-executive-order-to-accelerate-independent-oversight-and-advance-the-creation-of-an-ai-kill-switch>).

Zu 5. Schwerwiegende Vorfälle müssen zukünftig unverzüglich gemeldet und unabhängig untersucht werden. Um Transparenz herzustellen und Vertrauen zu stärken, sollen wesentliche Erkenntnisse und Gegenmaßnahmen grundsätzlich veröffentlicht werden. Anbieter von entsprechenden Systemen müssen zukünftig nachweisen, dass ihre Systeme sicher sind und für Schäden haften, insbesondere wenn diese durch

vorsätzliche oder fahrlässige Verstöße gegen Sicherheitsanforderungen oder unzureichende Schutzvorkehrungen verursacht wurden.

Zu 6. Wirksame globale Sicherheitsstandards können nur entstehen, wenn neben der Europäischen Union insbesondere die USA und China als führende KI-Entwicklungsstaaten eingebunden werden. Die Bundesregierung soll deshalb gemeinsam mit europäischen und gleichgesinnten Partnern wie Japan, Südkorea und Kanada gezielt Gespräche mit den USA, China und weiteren relevanten Staaten aufnehmen und konkrete Vereinbarungen über gemeinsame Sicherheitstests, den Austausch über schwerwiegende Vorfälle und Mindeststandards für menschliche Kontrolle vorantreiben. Perspektivisch muss daraus eine dauerhaft institutionalisierte internationale Aufsicht entstehen

Vorabfassung – wird durch die lektorierte Fassung ersetzt.