Whitepaper | 2. November 2026

Frontier-KI-Sicherheit: Wirksame Governance-Strukturen etablieren, internationalen Prüfrahmen aufbauen

Executive Summary

Neuartige Risiken von Frontier-KI: Die Entwicklung von immer fortschrittlicheren KI-Modellen bietet beträchtliche sozioökonomische Chancen, birgt jedoch auch völlig neuartige Risiken. Vorfälle in Test- und Einsatz-Umgebungen zeigen, dass fortgeschrittene agentische KI-Systeme zu komplexen Verhaltensweisen fähig sind, welche schwer antizipierbar sind und kritische Sicherheitsrisiken bergen. Angesichts der spezifischen Kombination von hochautonomen Verhaltensweisen und neuesten Fähigkeiten von Frontier-Modellen kommen bisherige KI-Governance-Ansätze an ihre Grenzen. Freiwillige Selbstverpflichtungen der Hersteller reichen nicht aus. Die neuen Risiken sind eine Gefahr für die nationale Sicherheit und werden bisher weder in Deutschland noch weltweit angemessen rechtlich und institutionell adressiert.

AISI Deutschland zum Kompetenzzentrum für Frontier-KI und nationale KI-bezogene Sicherheit ausbauen: Deutschland muss systemische KI-Risiken eigenständig bewerten und begrenzen können. Das ist auch eine Frage der digitalen Souveränität. Eine robuste KI-Sicherheitsarchitektur sollte dafür zwei Komponenten vereinen: 1) Kompetenz zur lebenszyklusbegleitenden, unabhängigen Evaluation und Prüfung von Frontier-Modellen. 2) Sicherheits-, domänen- und prüfspezifische Kompetenz, sobald Frontier-KI auf kritische Infrastrukturen, Industrieanlagen, Fahrzeuge oder andere Systeme der physischen Welt trifft. Isolierte KI-Kompetenz reicht hier nicht aus. Der Aufbau des AISI sollte deshalb von Beginn an gemeinsam mit unabhängigen Prüforganisationen erfolgen und deren über Jahrzehnte aufgebaute Sicherheits-, Domänen- und Prüfkompetenz systematisch einbinden.

Mehrstufige Governance – nationalen und internationalen Sicherheitsrahmen schaffen: Weltweit gibt es derzeit ganz unterschiedliche Regulierungsansätze von freiwilliger Selbstregulierung bis hin zu verbindlichen gesetzlichen Vorgaben (wie etwa EU AI Act oder auf US-Bundesstaaten-Ebene), und entstehen vielerorts nationale KI-Sicherheitsinstitute. Was fehlt ist ein international verbindlicher Rahmen – vergleichbar mit nuklearer Sicherheit - für Frontier-KI-Sicherheit. Es braucht Maßnahmen auf deutscher, europäischer, EU+- und globaler Ebene und diese müssen ineinandergreifen – von nationalen Prüf- und Aufsichtsstrukturen über einen kohärenten und konsequent umgesetzten EU-Regulierungsrahmen so-

wie EU-Zusammenarbeit mit gleichgesinnten „Middle Powers“ bis hin zu international abgestimmten Sicherheitskriterien, Prüfmechanismen und Melde- sowie Aufsichtsstrukturen. So entsteht ein mehrstufiger Sicherheitsrahmen, der die globalen und lokalen Auswirkungen der mächtigsten KI-Modelle und KI-Systeme adäquater adressiert.

Inhalt

Warum Frontier-KI geprüft werden muss 3

Wie Frontier-KI-Prüfung organisiert werden kann: Vier Governance-Ebenen, ein Sicherheitsrahmen 5

1. Digitale Souveränität: AISI Deutschland als leistungsfähigen Kern etablieren 5
2. EU: Zukunftsfester Rechtsrahmen mit unabhängigen Prüfkapazitäten 8
3. EU+: Coalition of the Willing & Partnerschaften für KI-Sicherheit schmieden 9
4. International & Global: Mindeststandards und Prüfregime vorantreiben 10

Warum Frontier-KI geprüft werden muss

Qualitativ neuartige KI-Risiken: Frontier-KI – die leistungsfähigsten Modelle der neuesten Generation und die auf ihnen aufbauenden Systeme – entwickelt neuartige und mächtige Fähigkeiten. Agentische Frontier-KI-Systeme können eigenständig Handlungspläne erstellen, digitale Werkzeuge wie APIs nutzen oder Code selbstständig generieren und nicht zuletzt mit anderen agentischen KI-Systemen interagieren. Mittelbar und, je nach Kontext, auch unmittelbar können Frontier-KI-Systeme Effekte in der physischen Welt erzielen. Speziell die **Kombination aus stark erhöhtem Autonomiegrad und dynamisch wachsenden Fähigkeiten von Frontier-Modellen** ergibt neue Potenziale, erhöht zugleich aber Missbrauchspotenziale und Sicherheitsrisiken. Mit wachsenden KI-Fähigkeiten müssen daher auch interne Schutzmaßnahmen, unabhängige Evaluationen und Prüfungen sowie staatliche und internationale Kontrollmechanismen Schritt halten. Die folgenden, teils in Beziehung zueinander stehenden Risikocluster sind bei Frontier-KI aktuell besonders im Fokus:

- **Cyberangriffe**: Böswillige Akteure können Frontier-KI zur gezielten Schwachstellensuche und automatischen Exploit-Code-Generierung nutzen sowie – u.a. basierend auf dem Schwarm-Verhalten von agentischen Systemen – komplexe Phishing- und Angriffskampagnen skalieren. Ausgehend von Kontrollverlusten können Frontier-Systeme solche Angriffe auch autonom, ohne menschliche Anweisung oder bösartigen Willen durchführen.

- **CBRN-Risiken** (chemisch, biologisch, radiologisch und nuklear): Fortgeschrittene Modelle können als Know-how-Multiplikator wirken und gezielt zur Entwicklung von Waffen und Materialien in diesem Bereich genutzt werden. Solche Vorgänge können dabei auch autonom durch Frontier-KI initiiert werden.

- **Kontrollverlust**: Agentische Systeme können zur Zielerreichung Schutzmechanismen gezielt umgehen oder Verhaltensweisen zeigen, die funktional auf strategische Täuschung oder die Unterwanderung menschlicher Kontrolle hinauslaufen. Darüber hinaus besteht die eminente Gefahr, dass Frontier-Systeme unerwünschte Wege zur Zielerreichung einschlagen, bei denen immense Kollateralschäden nicht ausgeschlossen sind.

- **Gesellschaftliche Risiken**: Skalierte personalisierte Desinformation und Manipulation können gesellschaftliches Vertrauen und demokratische Prozesse entscheidend beeinträchtigen. Auch massive Auswirkungen autonomer KI-Systeme etwa auf den Arbeitsmarkt fallen in dieses Risikocluster.

- **Exponentielles Potenzial und Selbstverbesserung**: Schnelle Fähigkeitszuwächse und strategische Täuschung können bestehende Schutzmechanismen überfordern. Mögliche

Technologiesprünge durch selbstverbessernde Systeme und rekursive Entwicklungsansätze können diese Risiken entscheidend verschärfen.

Freiwillige Selbstregulierungs-Verpflichtungen der Hersteller reichen bei Frontier-KI nicht aus: Auch die technische Sicherheitsevaluation wird anspruchsvoller. Reine Blackbox-Modellevaluationen bilden das Verhalten eines Modells im realen Einsatz nur begrenzt ab. Fortgeschrittene Systeme können Testsituationen erkennen und ihr Verhalten gezielt anpassen – durch sogenanntes Scheming, beispielsweise Alignment Faking oder Sandbagging, bei dem die Systeme ein Befolgen der Nutzerintention vortäuschen oder ihre wahren Fähigkeiten gezielt verschleiern. Prüfungen müssen daher nicht nur direkt beobachtbare Modelleigenschaften, sondern auch Robustheit, Schutzmechanismen, mögliche Umgehungsstrategien und das Verhalten unter realistischen Einsatzbedingungen wie auch die gezielte, unabhängige Betrachtung der Trainingsprozesse, etwa durch sogenannte „embedded evaluators", erfassen. Auf diese Weise können beispielsweise größere Schadenspotenziale, die sich bei Frontier-KI bereits vor der Produktivsetzung materialisieren können, gezielt adressiert werden. Gleiches gilt für potenziell unsichere Testumgebungen und -prozesse, die, wie die aktuellen Fälle zeigen, bereits signifikante Risikopotenziale bergen. Adäquaten Zugang vorausgesetzt, können unabhängige Prüfungen hier Grundlagen für notwendige, sicherheitsrelevante Weichenstellungen liefern. Frontier-KI ist deshalb nicht grundsätzlich unprüfbar. Vielmehr braucht es – komplementär zu den Sicherheitsbestrebungen der Hersteller – fachkundige und durch Domänenexpertise ergänzte unabhängige Prüfungen, geeignete Testumgebungen und weiterentwickelte Evaluationsmethoden, die jeweils den Stand der Technik abbilden. Prüforganisationen leisten hierzu gerne maßgebliche Beiträge, etwa im Bereich von KI-spezifischen Risikoevaluationen oder bei der unabhängigen Bewertung von Sicherheitsargumentationen.

Prüfkompetenz ist Teil von digitaler Souveränität: Deutschland und Europa müssen bei KI zweigleisig vorgehen; sie müssen eigene technologische Fähigkeiten entwickeln und zugleich eigenständige Prüfkompetenz aufbauen. Der Anspruch muss sein, insbesondere bei Industrial AI und Physical AI eigene Stärken auszubauen und international wettbewerbsfähig zu sein. Gleichzeitig ist offen, ob und wann Europa bei den leistungsfähigsten Frontier-Modellen mit den USA und China gleichziehen kann. Die europäische Wirtschaft und Gesellschaft werden deshalb auch künftig auf Modelle aus anderen Wirtschaftsräumen zurückgreifen. Digitale Souveränität bedeutet daher nicht nur technologische Unabhängigkeit in strategisch wichtigen Bereichen. Sie bedeutet auch, fremde Technologien eigenständig beurteilen und ihren sicheren Einsatz selbstbestimmt gestalten zu können. Deutschland und Europa müssen wissen, welche Fähigkeiten importierte Modelle besitzen, welche Risiken von ihnen ausgehen und unter welchen Bedingungen sie sicher in Industrie, kritischer Infrastruktur und Gesellschaft eingesetzt werden können. Eine eigenständige Prüfkompetenz ist neben z.B. eigener Entwicklung folglich eine wichtige Säule digitaler Souveränität.

Wie Frontier-KI-Prüfung organisiert werden kann: Vier Governance-Ebenen, ein Sicherheitsrahmen

1. Digitale Souveränität: AISI Deutschland als leistungsfähigen Kern etablieren

Eine zentrale Rolle kommt dem AI Safety and Security Institute (AISI) Deutschland als leistungsfähigen Kern der Sicherheitsarchitektur hierzulande und in Europa zu. Deutschland muss KI-Risiken für Sicherheit, Wirtschaft und Gesellschaft frühzeitig bewerten und darauf basierend handeln können. Das jüngst gestartete AISI Deutschland schafft dafür einen wichtigen institutionellen Anker und soll zeitnah auch international Ansprechpartner auf Augenhöhe werden.

Sicherheitstechnische Schwerpunkte des AISI Deutschland

Damit Deutschland vorausschauend handeln kann, sollte das AISI die im folgenden aufgeführen vier Fokusbereiche abdecken. Diese vier Fokusbereiche können bezüglich aller fünf zuvor genannten Risikocluster systematisch angewendet werden: Cyberangriffe, CBRN-Risiken, Kontrollverlust-Risiken, gesellschaftliche Risiken sowie exponentielles Potenzial und Selbstverbesserung. Dabei sollte zunächst ein Fokus auf die sich aktuell in zunehmendem Maße materialisierenden Kontrollverlust- und Cyberangriffsrisiken gesetzt werden. Dieser Fokus kann in der Folge an künftige Entwicklungen der Risikolandschaft angepasst werden.

- **Lagebild analysieren, inkl. Modellevaluationen**: Hier gilt es, Informationen über den aktuellen Zustand der Entwicklungen strukturiert zu erfassen und aufzubereiten. Auf (sozio-)technischer Ebene sollten hierzu Modellevaluationen durchgeführt, die Entwicklungsprozesse der Frontier-KI-Unternehmen verfolgt und die schon auftretenden Risiken und Vorfälle strukturiert beobachtet werden. Ergänzend dazu sollten die Risikotoleranz von Gesellschaft und Institutionen eingeschätzt und geopolitische Entwicklungen analysiert werden.

- **Entwicklungen evidenzbasiert prognostizieren**: Entwicklungen im Bereich Frontier-KI vollziehen sich so dynamisch, dass sie systematisch und fortlaufend prognostiziert werden müssen, um vorbereitet und handlungsfähig zu bleiben. Wird Deutschland hingegen von den Fähigkeiten und Auswirkungen eines neuen Modells überrascht, so drohen in den ersten Tagen und Wochen bereits große Schäden. Evidenzbasierte Prognosen lassen sich auf Basis des Lagebilds sowie empirisch beobachteter Fähigkeitsentwicklungen und Skalierungszusammenhänge erstellen.

- **Beherrschbarkeit aktueller und zukünftiger Risiken prüfen**: Basierend auf dem Lagebild und den evidenzbasierten Prognosen sollte weiterhin geprüft werden, ob Staat, Wirtschaft und Gesellschaft auf die Risiken vorbereitet sind. Eine solche Prüfung hat zwei Dimensionen: Es sollte zum einen technisch geprüft und erforscht werden, ob die technischen Maßnahmen, die aktuell für Frontier-KI-Modelle zur Risikominderung eingesetzt werden, auch in Zukunft wirksam sind. Dies ist besonders relevant, da es bereits überzeugende theoretische Argumente und experimentelle Beobachtungen gibt, die darauf hindeuten, dass aktuell angewendete Kontrollmaßnahmen und Trainingstechniken bei zunehmenden Modellfähigkeiten ihre Wirksamkeit verlieren. Zum anderen sollte geprüft werden, ob die Schutzmaßnahmen für Institutionen und Gesellschaft wirksam bleiben.

- **Entscheidungen und Handlungsoptionen aufbereiten**: Falls sich aus den vorangegangenen Analysen ergibt, dass Risiken mit dem aktuellen Stand der Mitigationsmaßnahmen und Vorbereitung nicht ausreichend beherrschbar sind, sollte erarbeitet und kommuniziert werden, wie diese Lücken geschlossen werden können. Zentrale Adressaten dieser Kommunikation sind insbesondere die Bundesregierung und der Nationale Sicherheitsrat sowie, abhängig vom jeweiligen Risikobereich, unter anderem zuständige Behörden, Sicherheitsorgane und Betreiber kritischer Infrastrukturen. Ziel sollte sein, relevante Entwicklungen frühzeitig in entscheidungsreife Optionen zu übersetzen und so die staatliche Handlungsfähigkeit auch bei einer sich schnell verändernden Risikolandschaft sicherzustellen.

AISI schrittweise zu einer eigenständigen Institution weiterentwickeln

In der Aufbauphase wird das AISI Deutschland gemeinsam von Bundesnetzagentur und Bundesamt für Sicherheit in der Informationstechnik (BSI) getragen. Diesen schnellen Aufbau innerhalb bestehender Strukturen gilt es zunächst zu unterstützen. Perspektivisch sollte das AISI jedoch zu einer **eigenständigen, dauerhaft handlungsfähigen Institution** mit klarem Mandat und angemessenen Ressourcen weiterentwickelt werden.

Bereits bei dieser institutionellen Weiterentwicklung sollte mitgedacht werden, wie die in Deutschland vorhandene unabhängige Prüf-, Sicherheits- und Domänenkompetenz dauerhaft eingebunden werden kann. **Eine Option wäre, das AISI perspektivisch als gemeinsame Gesellschaft unter Beteiligung staatlicher Akteure und unabhängiger Prüforganisationen auszugestalten**. Auf diese Weise könnten staatliche Verantwortung und Steuerungsfähigkeit mit hochspezialisierter operativer Expertise verbunden werden, ohne dass der Staat sämtliche Fähigkeiten selbst neu aufbauen muss. Mit ihrer Fachexpertise in der KI-Prüfung und ihrer branchenübergreifenden Expertise zur Prüfinfrastruktur und der Sicherheit technischer Systeme und Infrastrukturen können Prüforganisationen in allen vier Fokusbereichen unterstützen.

Deutschland kann hierfür auf Erfahrungen aus anderen sicherheitskritischen Bereichen zurückgreifen. Modelle aus der Reaktorsicherheit und der Mobilität zeigen unterschiedliche Wege, wie staatliche Aufgaben, unabhängige technische Expertise und institutionalisierte Prüfkompetenz miteinander verbunden werden können. **Welche konkrete Organisationsform für das AISI geeignet ist, sollte im Zuge seiner Verselbstständigung geprüft werden**.

Staatliche Verantwortung mit unabhängiger Prüfkompetenz verbinden: Deutschland kann bei der Organisation unabhängiger KI-Prüfung zudem auf bewährte Prinzipien seiner technischen Sicherheitsarchitektur zurückgreifen. Ein erstes Modell bietet das Typgenehmigungssystem im Fahrzeugbereich: Staatliche Genehmigungs-, Benennungs- und Aufsichtsfunktionen verbleiben bei der zuständigen Behörde, während qualifizierte Technische Dienste die hierfür erforderlichen technischen Prüfungen und Bewertungen durchführen. Dieses Prinzip erlaubt es, hochspezialisierte externe Prüfkompetenz zu nutzen, ohne die staatliche Entscheidungs- und Aufsichtshoheit aufzugeben.

Ein weitergehendes institutionelles Modell bietet die FSD Fahrzeugsystemdaten GmbH – als beliehene Zentrale Stelle nach dem deutschen Straßenverkehrsgesetz (StVG). Sie bündelt hochspezialisierte technische und methodische Kompetenz für die regelmäßige Fahrzeugüberwachung. Auf gesetzlicher Grundlage sind ihr klar definierte hoheitliche Befugnisse zur eigenverantwortlichen Wahrnehmung übertragen. Das Beispiel zeigt, dass spezialisierte technische Expertise unter klar bestimmten gesetzlichen Voraussetzungen auch institutionell in die Wahrnehmung staatlicher Aufgaben eingebunden werden kann. Ein weiteres institutionelles Vorbild bietet die Gesellschaft für Anlagen- und Reaktorsicherheit (GRS). Zu prüfen ist, welche Elemente dieser Modelle für Frontier-KI geeignet sind.

Das europäische Erfolgsmodell der Produktsicherheit auf Frontier-KI übertragen: Im Rahmen seines Mandats besteht für das AISI die Möglichkeit, Referenzmethoden sowie Evaluationsverfahren zu koordinieren, den erforderlichen Modellzugang zu unterstützen sowie Ergebnisse zusammenzuführen und für staatliche Aufsicht und internationale Zusammenarbeit nutzbar zu machen. Qualifizierte unabhängige Prüfstellen wiederum können auf dieser Grundlage die eigentlichen technischen Evaluationen und Sicherheitsprüfungen durchführen. Soweit externen Stellen darüber hinaus hoheitliche Befugnisse zur selbstständigen Aufgabenwahrnehmung übertragen werden sollen, bedürfte dies einer klaren gesetzlichen Grundlage, die Zweck, Umfang und Befugnisse sowie die staatliche Aufsicht bestimmt. Entscheidend ist das übergreifende Prinzip: Der Staat behält Verantwortung, Aufsicht und – soweit erforderlich – Entscheidungshoheit, muss aber nicht sämtliche operative KI-, Safety-, Domänen- und Prüfkompetenz selbst aufbauen.

2. EU: Zukunftsfester Rechtsrahmen mit unabhängigen Prüfkapazitäten

Auf EU-Ebene sollte der Umgang mit Frontier-KI vor dem Hintergrund der sich entwickelnden Risiko- und Bedrohungslage angepasst werden. Dies betrifft sowohl die Überprüfung und ggf. Weiterentwicklung bestehender Governance-Instrumente als auch den Erhalt eines kohärenten Rechtsrahmens und intensivierte Unterstützung europäischer Prüf- und Evaluationskapazitäten.

Bestehenden Governance-Rahmen für Frontier-KI überprüfen: Die bestehende Governance-Struktur muss auf ihre Effektivität hin überprüft und ggf. entsprechend nachgeschärft werden. Das AI Office sollte daher in Erwägung ziehen, auf eine Anpassung des GPAI Code of Practice an die geänderte Bedrohungs- und Erkenntnislage hinzuwirken, wo dieser nicht mehr dem State of the Art entspricht. Dies ist beispielsweise beim Thema Drittprüfungen der Fall, wo der Code of Practice für die riskantesten Modelle nur Anforderungen zu Modell-Evaluationen beinhaltet, während sich der State of the Art immer mehr zu sogenannten „embedded evaluations" entwickelt, bei denen das Prüfpersonal schon während des Trainings im vergleichbaren Umfang wie Mitarbeiter Zugriff auf Systeme, Trainingsparameter etc. hat.

KI-Rechtsrahmen strukturell stärken statt fragmentieren: Kontrolle über Frontier-KI zu behalten, ist eine politische und gesellschaftliche Querschnittsaufgabe. Daher muss der Rechtsrahmen für KI insgesamt horizontal ausgestaltet sein, zumal auch die jeweils aktuelle Frontier-KI nicht isoliert ohne Anwendungskontext betrachtet werden darf, da diese Modelle eher früher als später in sektoral regulierte Produktbereiche (z.B. Mobilität oder Medizintechnik) Einzug halten wird. Dagegen droht der KI-Rechtsrahmen derzeit vielmehr in einzelne branchenspezifische Sonderregeln zu zersplittern. Der rechtliche Rahmen für KI verliert dadurch zunehmend an Einheitlichkeit und Klarheit. Zudem kann er so nicht flexibel, zeitnah und zentral angepasst werden, wenn die Technologie sich weiterentwickelt.

EU Action Plan on Cybersecurity and Artificial Intelligence schnell umsetzen: Der Plan muss auf EU-Ebene zügig umgesetzt und der Ausbau von Prüfkapazitäten sowie Fachwissen für Frontier-KI konkret angegangen werden. Angesichts der enormen Mittel, die vor allem in den USA mobilisiert werden, um entsprechendes Fachpersonal zu gewinnen und externe Prüfungen von Frontier-KI zu ermöglichen, bedarf es einer starken politischen Bereitschaft, in der EU konkurrenzfähige Rahmenbedingungen zu schaffen. Dies ist dringend erforderlich, um nicht dauerhaft von der Expertise in Drittstaaten abhängig zu sein und talentiertes, hochqualifiziertes Spitzenpersonal in Europa zu halten.

Vorhandene Strukturen beim Ausbau externer Prüfkapazitäten nutzen: Auf bereits vorhandene Strukturen sollte zurückgegriffen werden, auch damit diese nicht erst langwierig aufgebaut werden müssen. Die EU hat eine bewährte Tradition im Hinblick auf Drittprüfungen zur Gewährleistung von Produktsicherheit. Auch wenn zunächst nicht vorgesehen sein sollte, im Rahmen des EU Action Plan on Cy-

bersecurity and Artificial Intelligence Evaluationen und Prüfungen als Konformitätsbewertung im Rechtssinne durchführen zu lassen, so drängt es sich dennoch auf, die in diesem Bereich insbesondere für die Prüfung von KI-Systemen und Cybersicherheit sowie allgemein für die Sicherheit kritischer Infrastrukturen und Anlagen vorhandene Expertise zu nutzen, um unabhängige Prüfmöglichkeiten schnellstmöglich hochzufahren.

3. EU+: Coalition of the Willing & Partnerschaften für KI-Sicherheit schmieden

Um den Gefahren von Frontier-KI wirksam zu begegnen, bleibt ein globaler Sicherheitsrahmen unter Einbeziehung aller zentralen Akteure das mittel- und langfristige Ziel. Auf dem Weg dorthin sollten Deutschland und die EU gemeinsam mit gleichgesinnten Partnern in einer KI-sicherheitsorientierten „Coalition of the Willing" vorangehen, die bereit ist, konkrete Maßnahmen zur Verringerung von KI-Sicherheitsrisiken umzusetzen. Eine Schlüsselrolle kommt dabei einer Gruppe von „Middle Powers" zu: technologisch fortgeschrittenen Demokratien mit leistungsfähigen Institutionen. Ihre Zusammenarbeit kann internationale Anschlussfähigkeit schaffen und den Weg zu einem globalen Konsens ebnen. Der EU Action Plan on Cybersecurity and Artificial Intelligence greift die Zusammenarbeit mit gleichgesinnten Partnern bereits auf. Dieser Ansatz sollte auf die Sicherheit von Frontier-KI ausgeweitet werden.

„Coalition of the Willing" für KI-Sicherheit aufbauen: Da abgestimmte Maßnahmen bei einer globalen Technologie wie KI wirksamer sind als ein europäischer Alleingang, sollten Deutschland und die EU mit gleichgesinnten Partnerstaaten wie z.B. dem Vereinigten Königreich, Kanada, Japan, Südkorea und Australien eine schlagkräftige Sicherheitsallianz aufbauen. Ziel bleiben internationale Standards und gemeinsame Maßnahmen, zugleich sollte kurzfristig pragmatisch mit jenen Staaten vorangegangen werden, die ein vergleichbares Risikoverständnis teilen und konkrete Schritte in der KI-Sicherheit mittragen wollen. Eine solche „Coalition of the Willing" kann zugleich die Grundlage dafür schaffen, international für weitergehende Vereinbarungen zu werben.

Gemeinsame „rote Linien" definieren und internationale Normungsarbeit vorantreiben: Die Partnerstaaten sollten Sicherheitsansätze abgleichen und gemeinsame Schwellenwerte für inakzeptable KI-Risiken definieren, ab denen Fähigkeiten oder Risiken fortgeschrittener KI-Systeme als nicht akzeptabel gelten. Bestehende Ansätze – etwa aus den Seoul Commitments oder der Zusammenarbeit der KI-Sicherheitsinstitute – gilt es hierfür weiter zu operationalisieren. Zugleich sollten die Partner anstreben, gemeinsame Sicherheitsanforderungen international zu verankern. Ein koordiniertes Auftreten erhöht die Chancen, international anschlussfähige Standards zu erreichen.

Gegenseitige Anerkennung und Interoperabilität von Evaluationsmethoden: Die KI-Sicherheitsinstitute der Partnerstaaten sollten ihre Prüfmethoden miteinander abstimmen

und gemeinsam anerkannte Kriterien und Verfahren, etwa Benchmarks und Evaluationsprotokolle, etablieren. So lässt sich die Zusammenarbeit von KI-Sicherheitsinstituten effizienter gestalten und es werden sowohl KI-Sicherheitsinstitute als auch Frontier-KI-Entwickler entlastet.

Gemeinsamer Zugang zu und Assessments von Frontier KI-Modellen: Die Partnerstaaten sollten gegenüber globalen Frontier-KI-Labs, die oftmals ihren Sitz in den USA haben, ein gemeinsames Verhandlungsmandat entwickeln, um KI-Sicherheitsinstituten der Middle-Power-Koalition angemessenen Zugang zu Frontier-Modellen, Frontier-Systemen, sicherheitsrelevanten Informationen sowie den für unabhängige Evaluationen erforderlichen Daten, Werkzeugen und Systemen zu ermöglichen. Hierfür kann insbesondere an die Zusammenarbeit des AISI mit weiteren KI-Sicherheitsinstituten angeknüpft werden, etwa im Rahmen des International Network for Advanced AI Measurement, Evaluation and Science (NAAIMES).

4. International & Global: Mindeststandards und Prüfregime vorantreiben

Für Frontier-KI braucht es international anschlussfähige Prüf- und Aufsichtsstrukturen, die die Einhaltung von Sicherheitsanforderungen überprüfen und eine wirksame globale Kontrolle ermöglichen. Zentral ist dabei die Unabhängigkeit der Prüfstellen, um eine Beeinflussung oder Verzerrung von Evaluationen und Prüfungen durch wirtschaftliche, organisatorische oder andere Interessen auszuschließen. Ergänzend braucht es eine internationale Aufsichtsstruktur – etwa unter dem Dach der Vereinten Nationen – die gemeinsame Sicherheitskriterien, „rote Linien" und Akkreditierungswege für Prüfstellen überwacht.

Internationale Sicherheitskriterien für Frontier-KI festlegen: Die Eindämmung von Gefahren besonders leistungsfähiger Frontier-KI muss global gedacht werden. Daher braucht es international abgestimmte Sicherheitskriterien für Frontier-KI, die verbindliche Tests während der Entwicklung und vor der Bereitstellung, unabhängige Prüfungen, gemeinsame Standards und den koordinierten Austausch über schwerwiegende Sicherheitsvorfälle umfassen. Internationale Institutionen und Kooperationsformate wie G7, OECD, Hiroshima AI Process, NAAIMES, AIQI (Consortium for Artificial Intelligence Quality Infrastructure) und GPAI (Global Partnership on AI) bieten hierfür eine Grundlage.

Internationale Anerkennung unabhängiger Prüfstellen: Um die Sicherheit besonders leistungsfähiger Frontier-KI zu gewährleisten, sind unabhängige Prüfungen bereits während der Entwicklung und vor der Bereitstellung erforderlich. Diese Prüfungen dürfen nicht allein durch die Entwickler selbst erfolgen, sondern müssen durch unabhängige, qualifizierte Prüforganisationen nach transparenten und international harmonisierten Anforderungen durchgeführt werden. So kann sichergestellt werden, dass Sicherheitsevaluationen objektiv, nachvollziehbar und frei von wirtschaftlichen oder organisatorischen Interessenkonflikten erfolgen.

Internationales Abkommen, das die Verpflichtung zur Einhaltung von „roten Linien" verbindlich verankert: Um besonders schwerwiegende Risiken durch Frontier-KI wirksam zu begrenzen, sollten international vereinbarte „rote Linien" in einem völkerrechtlich verbindlichen Abkommen verankert werden. Staaten sollten sich verpflichten, Entwicklung oder Bereitstellung besonders leistungsfähiger KI-Modelle auszusetzen, falls definierte Risikoschwellen überschritten werden oder zentrale Sicherheitsanforderungen nicht hinreichend nachgewiesen werden können. Ein solches Abkommen sollte gemeinsame Kriterien, Verifikationsmechanismen und abgestimmte Maßnahmen bei Verstößen vorsehen.

Internationale Behördenstrukturen, die die Einhaltung von „roten Linien" überwachen: Ergänzend zu einem internationalen Abkommen braucht es eine internationale Aufsichtsstruktur – etwa unter dem Dach der Vereinten Nationen –, die die Einhaltung gemeinsamer Sicherheitskriterien und „roter Linien" für Frontier-KI überwacht und koordiniert. Sie sollte Verstöße untersuchen, relevante Informationen bündeln und bei schwerwiegenden Risiken abgestimmte internationale Maßnahmen ermöglichen. So können zentrale Sicherheitsanforderungen grenzüberschreitend nach einheitlichen Maßstäben durchgesetzt werden.

Globales Incident-Reporting-System: Das Etablieren eines internationalen Frühwarn- und Meldesystems für schwerwiegende KI-Sicherheitsvorfälle (analog zur Zivilluftfahrt oder zur IAEO bei nuklearen Vorfällen) bleibt sinnvoll, um – aufbauend auf Vorarbeiten etwa der OECD – weltweite Transparenz über neuartige Risikomuster zu schaffen.

***