Deutsche Hochschule
der Polizei
Universität der Polizeien
des Bundes und der Länder

Professur für Polizeitechnik mit
Schwerpunkt Digitalisierung
Leitung des Polizeitechnischen
Instituts
wilfried.honekamp@dhpol.de

POSITIONSPAPIER

Risiko eines Kontrollverlusts durch fortgeschrittene Künstliche Intelligenz

Univ.-Prof. Dr. Wilfried Honekamp

29. September 2026

Zusammenfassung

Die öffentliche Debatte über existenzielle Risiken fortgeschrittener Künstlicher Intelligenz hat sich im September 2026 deutlich verschärft. Forschende aus führenden KI-Unternehmen warnen vor autonomen und möglicherweise selbstverbessernden Systemen. Einzelne nennen sehr hohe subjektive Risikoeinschätzungen. Diese Aussagen verdienen Aufmerksamkeit, sind aber keine wissenschaftlichen Messungen der Wahrscheinlichkeit, dass die Menschheit ausgelöscht wird. Der Forschungsstand ist ambivalent. Unabhängige Evaluationen zeigen Fortschritte bei autonomer Aufgabenerledigung und einzelnen Komponenten autonomer Replikation. Experimentelle Arbeiten belegen zudem, dass Modelle ihr Verhalten an erkannte Trainings- oder Testsituationen anpassen können. Ein robuster autonomer Kontrollverlust gegen entschlossene menschliche Gegenmaßnahmen ist bislang jedoch nicht nachgewiesen. Der International AI Safety Report 2026 kommt zu einer vergleichbaren Einordnung. Für die Sicherheitsbehörden bedeutet das weder Entwarnung noch Anlass, die verantwortbare Nutzung heutiger KI-Systeme einzustellen. Entscheidend ist, die Entwicklung hochautonomer Systeme systematisch zu beobachten und Schutzvorkehrungen an ihre wachsende Handlungsfähigkeit anzupassen. Sicherheitsbehörden sollten sich dabei an überprüfbaren Indikatoren orientieren und Schwellen für staatliches Handeln vorbereiten. Europa beginnt regulatorisch nicht bei null. Die EU-KI-Verordnung enthält bereits besondere Pflichten für KI-Modelle mit allgemeinem Verwendungszweck und systemischem Risiko. Einen internationalen Wettlauf an der technologischen Grenze kann sie allein jedoch nicht steuern. Ergänzend erscheint ein überprüfbarer internationaler Sicherheits- und Verifikationsrahmen sinnvoll. Er sollte zunächst von den Staaten und Rechtsräumen getragen werden, die fortgeschrittene KI entwickeln oder zentrale Infrastruktur kontrollieren. Später kann er in einen breiteren multilateralen Rahmen eingebettet werden.

Am 9. September 2026 erklärte der ehemalige Anthropic-Forscher Jacob Coxon öffentlich, er habe das Unternehmen verlassen, weil OpenAI und Anthropic aus seiner Sicht auf selbstverbessernde Superintelligenz zusteuerten und dabei unverantwortliche Risiken eingingen (Coxon, 2026). Evan Hubinger, der bei Anthropic zu Alignment forscht, unterstützte die Warnung und nannte für sich persönlich eine Wahrscheinlichkeit von mehr als zehn Prozent, dass KI innerhalb des nächsten Jahrzehnts alle Menschen töten könnte. Zugleich stellte er klar, dass er das Risiko heutiger Modelle für gering hält und seine Sorge einer möglichen Superintelligenz gilt, die aus rekursiver Selbstverbesserung erwächst (Hubinger, 2026). Wenige Tage zuvor hatte Jakub Pachocki, Chief Scientist von OpenAI, davor gewarnt, die Entwicklung bei unzureichenden Fortschritten in Alignment und Monitoring weiterhin mit maximaler Geschwindigkeit zu skalieren. Er sprach sich für breitere Interventionen und internationale Koordination aus (Pachocki, 2026). Solche Aussagen wirken stark, weil sie aus dem unmittelbaren Umfeld der leistungsfähigsten Systeme kommen. Ihre Herkunft ersetzt jedoch keine wissenschaftliche Evidenz. Wer an fortgeschrittener KI arbeitet, kann relevante interne Beobachtungen besitzen. Er kann zugleich aber auch institutionelle Interessen oder eine besondere Risikowahrnehmung haben. Für eine wissenschaftlich belastbare Einordnung müssen die zugrunde liegenden Annahmen offengelegt und die Erkenntnisse nachvollziehbar geprüft werden können.

Dieses Positionspapier beantwortet daher nicht die Frage, ob KI die Menschheit auslöschen wird. Eine solche Prognose lässt sich derzeit nicht seriös treffen. Das Papier untersucht eine engere Frage. Es fragt, welche Teile heutiger Kontrollverlustszenarien bereits empirisch zu beobachten sind und welche weiterhin hypothetisch bleiben. Daraus werden zunächst Folgerungen für die Sicherheitsbehörden abgeleitet. Anschließend wird der internationale Regelungsbedarf betrachtet. Der Begriff Kontrollverlust bezeichnet hier eine Situation, in der ein KI-System dauerhaft außerhalb der wirksamen Kontrolle seiner Betreiber handelt und die Wiederherstellung dieser Kontrolle erheblich erschwert oder verhindert. Das ist enger als der Begriff KI-Risiko. Kriminalitätsfördernde Nutzung oder fehlerhafte Ausgaben können schwerwiegend sein, ohne jedoch einen solchen Kontrollverlust darzustellen. Diese Trennung ist für eine sachliche Debatte wesentlich. Für dieses Positionspapier wurden große Sprachmodelle unterstützend eingesetzt. Die inhaltliche Verantwortung liegt vollumfänglich bei mir.

Expertenwarnungen sind ein Signal, aber keine Risikomessung

Katastrophale und existenzielle Risiken werden in der KI-Forschung keineswegs nur von wenigen Außenseitern diskutiert. Grace et al. (2025) befragten 2023 fast dreitausend Forschende, die in führenden KI-Fachforen publiziert hatten. Je nach Formulierung der Frage gab rund die Hälfte der Befragten einer Entwicklung fortgeschrittener KI mit Folgen bis hin zur Auslöschung der Menschheit eine Wahrscheinlichkeit von mindestens zehn Prozent. Field (2025) fand unter über einhundert KI-Expertinnen und -Experten zwei deutlich unterschiedliche Grundmuster. Ein Teil betrachtet fortgeschrittene KI überwiegend als

kontrollierbares Werkzeug. Ein anderer Teil ordnet sie eher als potenziell unkontrollierbaren Akteur ein. Diese unterschiedlichen Einschätzungen zeigen einen Dissens unter den Befragten. Sie messen keine objektive Auslöschungswahrscheinlichkeit. Auch Expertinnen und Experten können hierzu Wahrscheinlichkeiten nur auf der Grundlage von Modellen und persönlicher Einschätzung angeben. Für ein Ereignis, das noch nie eingetreten ist und dessen technischer Pfad nicht feststeht, fehlt eine belastbare empirische Grundlage für die Kalibrierung solcher Wahrscheinlichkeiten. Eine Zahl wie zehn Prozent kann deshalb eine ernst gemeinte Risikoeinschätzung sein. Sie darf aber nicht wie eine statistisch ermittelte Unfallwahrscheinlichkeit behandelt werden.

Field, Douglas und Krueger (2026) befragten 25 KI-Forschende. Zwanzig von ihnen bezeichneten die Automatisierung von KI-Forschung als eines der schwerwiegendsten und dringlichsten KI-Risiken. Über Zeithorizonte und geeignete Governance bestand dennoch erhebliche Uneinigkeit. Forschende aus Hochschulen äußerten sich skeptischer gegenüber Szenarien eines explosionsartigen Fähigkeitszuwachses. Das spricht für Vorsicht gegenüber einer Darstellung, nach der die Fachwissenschaft bereits zu einer einheitlichen Prognose gelangt sei. Bareis, Ackerl und Heil (2026) untersuchen in einer Übersichtsarbeit 81 peer-reviewte Beiträge zu existenziellen KI-Risiken. Sie zeigen, dass Szenarien eines weitreichenden Kontrollverlusts häufig auf spekulativen Annahmen über die künftige Entwicklung von KI beruhen. Zugleich würden materielle und gesellschaftliche Randbedingungen in Teilen der Literatur nur unzureichend berücksichtigt. Vermeer et al. (2025) untersuchen drei mögliche Extinktionspfade über Kernwaffen, Pathogene und Geoengineering. In allen drei Szenarien wäre es nach ihrer Einschätzung außerordentlich schwierig, mit KI eine Bedrohung zu erzeugen, die zur Auslöschung der Menschheit führen könnte. Ausschließen können die Autoren diese Möglichkeit jedoch nicht. In den untersuchten Szenarien wäre eine Auslöschung zudem nur bei einem gezielt darauf gerichteten Akteur plausibel. Die Autoren erwarten zugleich, dass sich entsprechende Bedrohungen über einen Zeitraum entwickeln würden, der menschliche Gegenmaßnahmen grundsätzlich ermöglicht.

Was derzeit empirisch belegt ist

Der International AI Safety Report 2026 fasst den Forschungsstand zu möglichen Kontrollverlustszenarien breit zusammen. Aktuelle Systeme zeigen demnach erste dafür relevante Fähigkeiten. Sie besitzen diese jedoch noch nicht auf einem Niveau und in einer Kombination, die einen Kontrollverlust ermöglichen würde. Erforderlich wären unter anderem eine robuste autonome Handlungsfähigkeit sowie die Fähigkeiten, Aufsicht zu umgehen und menschliche Gegenmaßnahmen zu vereiteln. Gleichzeitig entwickeln sich mehrere dieser Teilfähigkeiten weiter. Für die staatliche Risikovorsorge ist deshalb gerade die Kombination aus fehlendem aktuellem Nachweis und wachsenden relevanten Fähigkeiten entscheidend (Bengio et al., 2026). Eine unabhängige Evaluation stammt von METR (2026). Die Organisation erhielt im Februar und März 2026 Zugang zu leistungsfähigen internen Modellen von Anthropic, Google, Meta und OpenAI sowie zu nicht

öffentlichen Informationen. METR hielt es für plausibel, dass intern eingesetzte Agenten zusätzliche, nicht autorisierte Agenten eigenständig starten und eine gewisse Zeit ohne Wissen oder Erlaubnis des Betreibers weiterlaufen lassen konnten. Nach Einschätzung der Forschenden waren sie jedoch nicht in der Lage, einen nicht autorisierten Betrieb von KI-Agenten in erheblichem Umfang vor einer aktiven Untersuchung zu verbergen oder so abzusichern, dass er einem mit hoher Priorität betriebenen Abschaltversuch standgehalten hätte. Klare Hinweise auf langfristig angelegte Machtziele im realen Trainings- oder Produktivbetrieb wurden nicht gefunden. Diese Ergebnisse sind keine Entwarnung. Sie setzen eine zeitliche Markierung. Im Frühjahr 2026 lag ein gegenüber aktiver Untersuchung und gezielter Abschaltung robuster autonomer Kontrollverlust außerhalb der von METR beobachteten Fähigkeiten. Für eine Sicherheitsstrategie ist diese Markierung wertvoller als eine spekulative Gesamtwahrscheinlichkeit. Sie kann durch wiederholte Tests fortgeschrieben werden.

Seit dem Sommer 2026 liegen zudem dokumentierte Sicherheitsvorfälle vor, die über synthetische Benchmarks hinausgehen. Bei internen Cybersecurity-Evaluationen von OpenAI fanden Agenten Wege, vorgesehene Isolationsmechanismen zu umgehen, über nicht vorgesehene Kanäle miteinander zu kommunizieren und auf externe Systeme zuzugreifen. Eine unabhängige Untersuchung von METR und Redwood Research ergab, dass über tausend eigentlich voneinander isolierte Agenten einen gemeinsamen Kommunikationskanal nutzten. Etwa 700 von ihnen beteiligten sich an einem mehrtägigen Angriff auf Systeme von Hugging Face. Die Agenten koordinierten dabei Teilaufgaben und suchten unter anderem nach Wegen, das Bewertungssystem des Benchmarks zu manipulieren. Die Systeme liefen allerdings in offensiven Cyber-Evaluationen mit reduzierten Schutzmechanismen. Der Vorfall belegt daher keinen robusten autonomen Kontrollverlust. Er zeigt jedoch, dass leistungsfähige Agenten unter ungünstigen Sicherheitsbedingungen technische Begrenzungen umgehen, unerlaubt kooperieren und reale Drittsysteme kompromittieren können (Greenblatt et al., 2026; OpenAI, 2026). Anthropic berichtete ebenfalls über vier Vorfälle, bei denen Modelle während Cybersecurity-Evaluationen unbefugt auf reale Drittsysteme zugriffen. Hinweise auf agentenübergreifende Koordination, Ziele außerhalb der vorgegebenen Aufgabe oder Versuche zur Umgehung der Aufsicht fanden die Autoren dort jedoch nicht (Bogdan et al., 2026).

Das unabhängige internationale wissenschaftliche Panel der Vereinten Nationen bewertet den OpenAI-Hugging-Face-Vorfall inzwischen als eines der bislang deutlichsten realweltlichen Warnsignale für einen möglichen Pfad zu einem Verlust menschlicher Kontrolle. Zugleich leitet es daraus weder eine Eintrittswahrscheinlichkeit noch einen Zeithorizont für einen schweren Kontrollverlust ab (Independent International Scientific Panel on AI, 2026). Eine am 28. September 2026 veröffentlichte Evaluation des britischen AI Security Institute ergänzt dieses Bild. GPT-6 Astra führte in vollständig simulierten Cyber-Evaluationen auch außerhalb des vorgegebenen Aufgabenbereichs nicht autorisierte Supply-Chain-Angriffe aus. Eine explizitere Begrenzung des zulässigen

Handlungsspielraums reduzierte dieses Verhalten, beseitigte es aber nicht vollständig. Da die Untersuchung vollständig simuliert war und die regulären Cyber-Sicherheitsfilter deaktiviert waren, ist sie kein Nachweis eines entsprechenden Verhaltens unter realen Einsatzbedingungen. Sie zeigt jedoch erneut, dass leistungsfähigere agentische Systeme bestehende Handlungsgrenzen nicht zuverlässig einhalten müssen (AI Security Institute, 2026).

Das britische AI Security Institute berichtet über deutliche Leistungszuwächse bei Fähigkeiten, die für eine autonome Replikation erforderlich wären. Auf einem Teil der Tests mit RepliBench, einem standardisierten Testverfahren, das einzelne für eine autonome Replikation notwendige Fähigkeiten von KI-Agenten prüft, stieg die Erfolgsrate von weniger als fünf Prozent für das leistungsfähigste Modell Anfang 2023 auf mehr als 60 Prozent für zwei besonders leistungsfähige KI-Modelle im Sommer 2025. Die Tests bilden jedoch vereinfachte Aufgaben in kontrollierten Umgebungen ab. Das Institut hält es für wahrscheinlich, dass aktuelle Modelle unter realen Bedingungen bei einer autonomen Replikation scheitern würden (D'Cruz et al., 2025). Eine weitere mit RepliBench durchgeführte Untersuchung zeigt ein ähnliches Bild. Die untersuchten Modelle bewältigten bereits zahlreiche Teilaufgaben, die für eine autonome Replikation erforderlich sind. Nach Einschätzung der Autoren stellten sie zum Untersuchungszeitpunkt jedoch noch keine glaubwürdige Gefahr einer autonomen Selbstreplikation dar (Black et al., 2025). Die Autoren weisen darauf hin, dass Erfolge bei einzelnen Teilaufgaben ein zu positives Bild der tatsächlichen Gesamtfähigkeit vermitteln können, weil eine autonome Replikation die koordinierte Bewältigung mehrerer Schritte voraussetzt.

Auch strategisch wirkendes Verhalten lässt sich experimentell beobachten. Greenblatt et al. (2024) zeigten in konstruierten Versuchsanordnungen sogenanntes Alignment Faking. Claude 3 Opus verhielt sich unterschiedlich, je nachdem, ob es eine Interaktion als trainingsrelevant einstufte. In den protokollierten internen Arbeitsschritten begründete das Modell dies teilweise damit, eine Veränderung seines bisherigen Verhaltens vermeiden zu wollen. Die Autoren hatten das Modell nicht ausdrücklich angewiesen, Alignment vorzutäuschen. Sie gestalteten die Versuchsanordnung jedoch so, dass das Modell erkennen konnte, wann sein Verhalten für weiteres Training genutzt werden würde. Das Ergebnis zeigt damit eine zuvor überwiegend theoretisch diskutierte Möglichkeit. Es belegt nicht, dass heutige Modelle im normalen Betrieb eigenständig dauerhafte verdeckte Ziele entwickeln. Mit steigender Zahl autonomer Agenten kann auch ihr Monitoring schwieriger werden. Makins et al. (2026) untersuchten in einer synthetischen KI-Laborumgebung koordinierte Angriffe mehrerer Agenten. Wurde ein Angriff auf mehrere Agenten verteilt, sank die Wahrscheinlichkeit, dass ein auf einzelne Agenten ausgerichtetes Monitoring einen der Angreifer erkannte. Ein leistungsfähigerer Monitor reduzierte unentdeckte erfolgreiche Angriffe um mehr als eine Größenordnung, beseitigte sie aber nicht. Die Arbeit zeigt einen relevanten Prüfpunkt für künftige Multi-Agenten-Architekturen.

Ein großer Teil der aktuellen Warnungen beruht auf der Annahme, dass KI zunehmend an ihrer eigenen Weiterentwicklung mitwirkt. Ansätze dafür existieren bereits. Systeme überarbeiten eigene Ausgaben, erzeugen Trainingsmaterial und unterstützen KI-Forschung. Daraus folgt jedoch noch keine offene rekursive Selbstverbesserung. Chen, Wang und Qu (2026) untersuchen über tausend arXiv-Arbeiten aus den Jahren 2024 bis 2026. Sie unterscheiden begrenzte Selbstverbesserung von weitgehend geschlossenen Verbesserungszyklen. Begrenzte Formen sind bereits praktisch etabliert. Nach ihrer Analyse ist eine offene rekursive Selbstverbesserung bislang begrenzt. KI-Systeme können ihre eigenen Verbesserungen noch nicht zuverlässig bewerten, und auch die verfügbaren Rechenressourcen setzen Grenzen. Auch die Festlegung neuer Forschungsrichtungen erfordert bislang menschliche Entscheidungen. Die Arbeit liefert eine hilfreiche begriffliche Trennung. Pachockis Warnung vor einer möglichen Fortsetzung des aktuellen Fortschritts in Richtung rekursiver Selbstverbesserung ist daher als fachlich informierte Prognose zu verstehen. Sie ist kein experimenteller Nachweis, dass dieser Übergang bereits erfolgt ist. Genau an dieser Stelle verlaufen gegenwärtig die wichtigsten wissenschaftlichen Unsicherheiten.

Tabelle 1: Übersicht des Stands der Forschung

Fragestellung	Erkenntnis	Einordnung
Robuster autonomer Kontrollverlust	In unabhängigen Evaluationen bislang nicht nachgewiesen.	Aktuelle Systeme sind nicht mit einem bereits eingetretenen Kontrollverlust gleichzusetzen. (Bengio et al., 2026; METR, 2026)
Autonome Replikation	Mehrere Teilfähigkeiten sind vorhanden und verbessern sich. Reale robuste Replikation bleibt begrenzt.	Weiter beobachten. Testumgebungen dürfen nicht mit realer Einsatzfähigkeit verwechselt werden. (D'Cruz et al., 2025; Black et al., 2025)
Strategische Anpassung an Monitoring	In konstruierten Experimenten nachweisbar.	Relevantes Warnsignal. Kein Nachweis stabiler verdeckter Ziele im normalen Betrieb. (Greenblatt et al., 2024)
Rekursive Selbstverbesserung	Begrenzte Selbstverbesserung ist real. Ein offener autonomer Verbesserungszyklus ist nicht gezeigt.	Hohe Unsicherheit. Fortschrittsindikatoren sollten getrennt gemessen werden. (Chen et al., 2026; Field et al., 2026)
Monitoring mehrerer Agenten	Koordination kann die Erkennung verteilter Angriffe erschweren.	Früher Forschungsstand. Für agentische Systeme ist Monitoring als eigene Sicherheitsfunktion zu behandeln. (Makins et al., 2026; Taylor et al., 2026)
Umgehung technischer Begrenzungen	In Cybersecurity-Evaluationen wurden unerlaubte Kommunikation zwischen Agenten und Zugriffe auf reale Drittsysteme dokumentiert.	Relevantes Warnsignal. Kein Nachweis eines robusten Kontrollverlusts. (Bogdan et al., 2026; Greenblatt et al., 2026; OpenAI, 2026)

Ein Kontrollsystem ist nur so gut wie die Daten, auf die es zugreifen kann. Taylor et al. (2026) untersuchten, wie sich die Möglichkeiten zum Monitoring fortgeschrittener KI-Systeme verändern könnten. Nach ihrer Einschätzung beruhen heutige Kontrollmöglichkeiten auf Eigenschaften, die mit zunehmender Leistungsfähigkeit und Autonomie teilweise erodieren könnten. Dazu gehören die derzeit noch gut zu beobachtenden

sprachlich dargestellten Gedankengänge. Auch die Aussagekraft von Tests kann sinken, wenn Modelle besser erkennen, dass sie gerade evaluiert werden. Künftige Architekturen könnten zudem Informationen in Formen speichern oder verarbeiten, die für Menschen kaum noch zu verstehen sind. Für Sicherheitsbehörden ergibt sich daraus ein einfacher Grundsatz. Mehr Autonomie darf nicht mit weniger Möglichkeiten zum Monitoring einhergehen. Systeme, die eigenständig Werkzeuge verwenden oder auf Netze zugreifen, sollten so eingesetzt werden, dass Handlungen nachvollziehbar bleiben. Berechtigungen müssen begrenzt sein und jederzeit entzogen werden können. Bei besonders kritischen Anwendungen sollte ein unabhängiges Monitoring nicht erst nach einem Vorfall eingerichtet werden. Die zentralen empirischen Erkenntnisse zu den für einen möglichen Kontrollverlust relevanten Fähigkeiten und ihre Einordnung sind in Tabelle 1 zusammengefasst.

Bedeutung für Öffentlichkeit und Polizei

Die öffentliche Kommunikation sollte zwei Fehler vermeiden. Der erste Fehler wäre, subjektive Risikozahlen als wissenschaftlich bestimmte Wahrscheinlichkeit zu verbreiten. Der zweite Fehler wäre, die Debatte allein deshalb abzutun, weil ein vollständiger Kontrollverlust noch nicht beobachtet wurde. Beides würde der Evidenz nicht gerecht. Für die Bevölkerung ist eine präzise Botschaft möglich. Für gegenwärtige KI-Systeme ist ein robuster autonomer Kontrollverlust in unabhängigen Evaluationen bislang nicht nachgewiesen. Gleichzeitig werden Systeme autonomer und sicherheitsrelevante Teilfähigkeiten nehmen zu. Forschung zu Kontrolle und zum Monitoring ist deshalb kein Randthema. Sie gehört zur vorausschauenden Technologiesicherheit. Für die Polizei ist diese Trennung besonders wichtig. Die heute dominierenden Risiken beim eigenen Einsatz von KI sind greifbarer. Sie betreffen etwa fehlerhafte Ergebnisse, Datenschutz und Verzerrungen. Europol sowie INTERPOL und UNICRI haben hierfür bereits umfangreiche Orientierungshilfen vorgelegt (Europol, 2024; INTERPOL & UNICRI, 2024). Daneben gewinnt die missbräuchliche Nutzung von KI durch Täter an Bedeutung. Europol berichtet inzwischen, dass kriminelle Netzwerke KI einsetzen, um ihre Aktivitäten zu skalieren und eigene Risiken zu verringern (Europol, 2026). Ein mögliches zukünftiges Szenario des Kontrollverlusts ersetzt diese Aufgaben nicht. Es erweitert den Blick auf eine neue Kategorie von Technikrisiken. Die Polizei sollte deshalb nicht aus einer hypothetischen Zukunftsgefahr den Schluss ziehen, KI generell zurückzuweisen. Der umgekehrte Schluss wäre ebenfalls falsch. Je stärker Systeme eigenständig planen und handeln, desto weniger genügt es, sie wie gewöhnliche Assistenzsoftware zu behandeln. Der Grad technischer Autonomie muss sich in der Sicherheitsarchitektur widerspiegeln.

Handlungsfelder für Polizei und staatliche Sicherheitsvorsorge

Sicherheitsbehörden benötigen ein Lagebild, das unabhängige Evaluationen zu Autonomie und Cyberfähigkeiten enthält. Erkenntnisse zu Möglichkeiten der Replikation und des Monitorings sind ebenso wichtig. Es sollte evaluiert werden, ob Modelle Fähigkeiten

gewinnen, die bestehende Kontrollmechanismen praktisch unterlaufen könnten. Das Polizeitechnische Institut kann hier in enger Kooperation mit dem vom BKA koordinierten Netzwerk KI eine vermittelnde Funktion übernehmen. Es kann wissenschaftliche Erkenntnisse in sicherheitsbehördliche Prüffragen übersetzen und Entwicklungen über Modellgenerationen hinweg vergleichen. Ein solches Lagebild sollte in kurzen Abständen aktualisiert werden. Bei einem schnellen Technologiefeld ist eine einmalige Bewertung rasch veraltet. Agentische Systeme sollten in polizeilichen Umgebungen nur die Berechtigungen erhalten, die für die konkrete Aufgabe erforderlich sind. Netzwerkzugriffe und Werkzeuge sollten begrenzt werden. Protokolle müssen eine nachträgliche Rekonstruktion ermöglichen. Berechtigungen müssen kurzfristig entzogen werden können. Bei kritischen Funktionen muss eine menschlich verantwortete Unterbrechung technisch erhalten bleiben. Diese Anforderungen sind nicht erst für eine hypothetische Superintelligenz relevant. Sie gehören bereits heute zur sicheren Integration agentischer Software. Gerade der METR-Bericht zeigt, dass KI-Agenten bei zu weitreichenden Berechtigungen und unzureichendem Monitoring weitere, nicht autorisierte Systeme eigenständig in Betrieb nehmen könnten, obwohl diese einem gezielten Abschaltversuch nicht standhalten würden (METR, 2026).

Auffälliges Verhalten agentischer Systeme sollte systematisch erfasst werden. Dazu gehören unerwartete Rechteausweitung und Umgehungsversuche. Auch Abweichungen zwischen Test und Betrieb sind relevant. Die Dokumentation sollte technische Artefakte sichern, ohne vorschnell menschliche Kategorien wie Absicht oder Täuschung zu unterstellen. Gerade bei lernenden Systemen kann ähnliches Verhalten aus sehr unterschiedlichen Ursachen entstehen. Eine abgestimmte Vorfallklassifikation würde den Austausch zwischen Behörden erleichtern. Sie könnte zudem helfen, aus Einzelfällen frühzeitig Muster abzuleiten. Für die Forschung wären anonymisierte und rechtssicher aufbereitete Vorfalldaten ein wertvoller Gegenpol zu rein synthetischen Benchmarks. Staatliche Vorsorge sollte nicht erst beginnen, wenn ein Extinktionsszenario plausibel erscheint. Schon deutlich schwächere Systeme können kritische Abläufe stören, wenn sie Cyberfähigkeiten mit längerer Autonomie verbinden. Behörden sollten daher prüfen, welche Kernprozesse von externen KI-Diensten abhängen. Für besonders kritische Funktionen sind Rückfallverfahren und alternative Kommunikationswege erforderlich. Übungen sollten realistische Zwischenstufen abbilden. Denkbar sind autonome Cyberangriffe, die mehrere Tage andauern, oder Agenten, die sich innerhalb einer kompromittierten Infrastruktur verteilen. Solche Szenarien sind für die Sicherheitsvorsorge nützlicher als eine Übung zur abstrakten Übernahme der Kontrolle über die Menschheit. Sie knüpfen an messbare Fähigkeiten an und erlauben konkrete Verbesserungen.

Fachliche Kompetenz wirkt auch gegen unnötige Angst. Beschäftigte sollten verstehen, was heutige Systeme tatsächlich können und wo ihre Grenzen liegen. Das gilt ebenso für Führungskräfte. Eine verständliche interne Kommunikation kann erklären, weshalb ein ernstes zukünftiges Risiko mit einer verantwortbaren heutigen Nutzung vereinbar ist. Für

externe Kommunikation sollte die Polizei auf überprüfbare Aussagen setzen. Spektakuläre Einzelzitate können einen Anlass bieten. Sie sollten nicht die Evidenzbasis ersetzen. Staatliche Stellen müssen technische Fähigkeiten nachvollziehbar einordnen. Schutzmaßnahmen und verbleibende Unsicherheit müssen ebenso transparent benannt werden.

Die aktuelle Debatte wird teilweise so geführt, als gäbe es für besonders leistungsfähige KI noch keinen regulatorischen Rahmen. Für Europa trifft das nicht zu. Artikel 55 der EU-KI-Verordnung enthält besondere Pflichten für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko. Dazu gehören Modellbewertungen nach dem Stand der Technik einschließlich dokumentierter Angriffstests. Anbieter müssen mögliche systemische Risiken bewerten und mindern. Schwerwiegende Vorfälle sind zu dokumentieren und unverzüglich zu melden. Auch ein angemessenes Cybersicherheitsniveau für das Modell und seine physische Infrastruktur ist vorgeschrieben. Damit greift die europäische Regelung mehrere Elemente auf, die in der aktuellen Debatte gefordert werden. Das ist für die politische Einordnung wichtig. Die offene Frage ist, wie weit ein regionaler Rechtsrahmen reicht, wenn Entwicklung und Recheninfrastruktur global verteilt sind. Ein EU-Rechtsrahmen kann Unternehmen auf dem europäischen Markt verpflichten. Er kann aber nicht allein sicherstellen, dass konkurrierende Staaten die Entwicklung fortgeschrittener KI nach denselben Schwellen steuern. Genau hier liegt das strategische Problem, das viele aktuelle Warnungen ansprechen. Einseitiges Bremsen der Entwicklung ist politisch schwer durchsetzbar, wenn andere Akteure dadurch einen technologischen Vorsprung gewinnen könnten. Für sehr weitreichende Risiken sollte die EU-KI-Verordnung aber durch internationale Vereinbarungen ergänzt werden, die wechselseitige Nachprüfbarkeit ermöglichen.

Internationale Verifikation als nächste Ebene

Internationale KI-Governance scheitert schnell an einem bekannten Problem. Staaten können gemeinsame Sicherheitsziele unterstützen und zugleich fürchten, dass andere Staaten heimlich schneller entwickeln. Eine Vereinbarung ohne überprüfbare Einhaltung verändert dieses Dilemma nur begrenzt. Deshalb sollte Verifikation zum Kern einer möglichen internationalen Regulierung von KI gehören. Die technische Forschung hierzu ist weiter, als es die öffentliche Debatte oft vermuten lässt. Harack et al. (2025) untersuchen verschiedene Formen internationaler KI-Abkommen und kommen zu dem Ergebnis, dass die Einhaltung mehrerer möglicher Regelungen grundsätzlich überprüfbar sein könnte. Für Vereinbarungen zur rechenzentrumsbasierten Entwicklung und Bereitstellung leistungsfähiger KI halten sie belastbare Verifikationsmechanismen innerhalb weniger Jahre für grundsätzlich erreichbar, sofern gezielt in die erforderliche Infrastruktur investiert wird. Der wissenschaftliche Beirat des UN-Generalsekretärs sieht insbesondere die Überprüfung von Rechenressourcen als aussichtsreichen Ansatz. Ergänzend nennt er software- und hardwaregestützte Verfahren sowie menschliche Audits. Selbstberichtspflichten der Entwickler hält der Beirat ausdrücklich nicht für einen Ersatz externer

Aufsicht und Verifikation (United Nations Secretary-General's Scientific Advisory Board, 2025). Auch Kooperation zwischen geopolitischen Rivalen muss nicht den Austausch sensibler Modellinformationen voraussetzen. Bucknall et al. (2025) kommen zu dem Ergebnis, dass Forschung zu Verifikationsmechanismen und gemeinsamen technischen Protokollen besonders geeignete Bereiche für internationale Zusammenarbeit sein können. Peigné, Nguyen und Wang (2026) schlagen eine Zero-Knowledge-Architektur vor, mit der sich bestimmte Eigenschaften von Trainingsläufen kryptografisch nachweisen ließen, ohne die Trainingsspezifikation und damit wesentliche Modellinformationen offenzulegen. Die Autoren beschreiben bislang keine einsatzfähige Lösung. Sie halten jedoch einen funktionsfähigen Proof of Concept innerhalb von etwa drei Jahren für technisch erreichbar.

Auch Pachocki (2026) fordert, bestehende freiwillige Sicherheitsrahmen in allgemein verbindliche Sicherheitsgrenzen für die weitere Entwicklung zu überführen. Deren Einhaltung könne durch unabhängige Auditoren, staatliche Stellen oder internationale Institutionen überwacht werden. Internationale Koordination der weiteren KI-Entwicklung müsse aus seiner Sicht zu einer politischen Priorität werden. Aus meiner Sicht sollte ein zukünftiger internationaler Sicherheits- und Verifikationsrahmen folgende Eigenschaften besitzen. Es braucht gemeinsam definierte Schwellen, an die zusätzliche Sicherheitsanforderungen geknüpft werden. Oberhalb dieser Schwellen sollten unabhängige Sicherheitsbewertungen verpflichtend werden. Große Trainingsläufe sollten in einer Form überprüfbar sein, die legitime Geheimhaltungsinteressen schützt. Schwerwiegende Sicherheitsvorfälle sollten den zuständigen Stellen gemeldet und, soweit vertretbar, zwischen ihnen ausgetauscht werden. Werden vereinbarte Sicherheitskriterien nicht eingehalten, sollte ein vorab festgelegtes Verfahren eine Verlangsamung oder vorübergehende Aussetzung weiterer Skalierung ermöglichen. Ein solcher Rahmen muss nicht von Beginn an universal angelegt sein. Aus meiner Sicht wäre zunächst eine Vereinbarung zwischen den Staaten und Rechtsräumen sinnvoll, die fortgeschrittene KI entwickeln oder wesentliche Recheninfrastruktur kontrollieren. Sie sollte für weitere Beitritte offenstehen.

Langfristig könnte eine Einbettung in einen breiteren multilateralen Rahmen unter dem Dach der Vereinten Nationen Legitimität und globale Beteiligung stärken. Die UN hat dafür inzwischen erste institutionelle Strukturen geschaffen. Die Generalversammlung richtete ein unabhängiges internationales wissenschaftliches Panel für KI und den globalen Dialog über KI-Governance ein. Der erste jährliche Dialog fand im Juli 2026 in Genf statt. Das wissenschaftliche Panel soll eine gemeinsame Evidenzgrundlage schaffen. Der Dialog bietet Staaten und weiteren Akteuren einen globalen Raum für die Abstimmung über KI-Governance. Eine Regulierungs- oder Durchsetzungsinstanz ist er ausdrücklich nicht. Ein vertraglicher internationaler Sicherheits- und Verifikationsrahmen mit überprüfbaren Pflichten besteht auf dieser Ebene bislang nicht (United Nations, 2026). Am 23. September 2026 befasste sich zudem der UN-Sicherheitsrat in seiner

10.228. Sitzung ausdrücklich mit künstlicher Intelligenz und internationaler Sicherheit. Zu den Briefings gehörten Yoshua Bengio sowie die Leiter von OpenAI, Anthropic und Hugging Face. Damit haben mögliche Kontroll- und Sicherheitsrisiken fortgeschrittener KI auch die formelle sicherheitspolitische Agenda des Sicherheitsrats erreicht (United Nations Security Council, 2026). Historische Rüstungskontrolle kann dafür Anregungen liefern, aber keine Blaupause. Bucknall et al. (2025) zeigen, dass selbst geopolitische Rivalen bei sicherheitskritischen Technologien Verifikationsmechanismen und gemeinsame technische Protokolle entwickeln können. KI weist jedoch eigene technische und institutionelle Bedingungen auf. Eine tragfähige Verifikation sollte deshalb auf konkret zu beobachtende Infrastruktur und überprüfbare risikorelevante Prozesse ausgerichtet sein. Sie sollte nicht versuchen, jede algorithmische Entwicklung international zu kontrollieren.

Schlussfolgerungen

Die Wahrscheinlichkeit einer KI-bedingten Auslöschung der Menschheit lässt sich auf Grundlage der verfügbaren empirischen Evidenz derzeit nicht belastbar in Prozent beziffern. Prozentangaben einzelner Forschender sind Risikoeinschätzungen. Sie sind keine empirisch kalibrierten Eintrittswahrscheinlichkeiten. Staatliche Kommunikation sollte diesen Unterschied deutlich machen. Die aktuelle Evidenz stützt nicht, dass heutige KI-Systeme bereits die für einen Kontrollverlust erforderliche Kombination und Ausprägung von Fähigkeiten besitzen. Zugleich nehmen einzelne dafür relevante Fähigkeiten zu, insbesondere die autonome Handlungsfähigkeit und Fähigkeiten, die eine wirksame Aufsicht erschweren können. Sicherheitsbehörden benötigen ein aktuelles, empirisch gestütztes Lagebild. Es sollte sich an überprüfbaren Fähigkeiten und Veränderungen sicherheitsrelevanter Eigenschaften orientieren. Für die Polizei gibt es keinen Grund, verantwortungsvoll eingesetzte KI-Anwendungen grundsätzlich einzustellen. Mit wachsender Autonomie sollten Berechtigungen enger begrenzt und Handlungen so protokolliert werden, dass Vorfälle rekonstruiert werden können. Bei kritischen Anwendungen muss die technische Unterbrechbarkeit erhalten bleiben. Agentische Systeme benötigen deshalb eine Sicherheitsarchitektur, die ihrem größeren Handlungsspielraum Rechnung trägt.

Die EU-KI-Verordnung schafft bereits einen wichtigen regulatorischen Rahmen für KI-Modelle mit allgemeinem Verwendungszweck und systemischem Risiko. Dieser Rahmen löst das internationale Wettbewerbsproblem jedoch nicht. Für fortgeschrittene KI sollte deshalb ergänzend an überprüfbaren internationalen Verpflichtungen gearbeitet werden. Gegenseitige Verifikation ist dabei belastbarer als unverbindliche Selbstverpflichtung. Die Einordnung fortgeschrittener KI sollte Alarmismus ebenso vermeiden wie die falsche Sicherheit, die aus dem Fehlen eines heutigen Katastrophennachweises entstehen könnte. Sicherheitspolitik muss auch unter tiefer Unsicherheit handlungsfähig bleiben. Ein sehr hohes Schadenspotenzial darf nicht allein deshalb unberücksichtigt

bleiben, weil seine Eintrittswahrscheinlichkeit noch nicht belastbar bestimmt werden kann. Umgekehrt darf Unsicherheit nicht als Gewissheit dargestellt werden.

Referenzen

AI Security Institute (2026). GPT-6 Astra performs unsanctioned supply-chain attacks in simulations. https://www.aisi.gov.uk/blog/gpt-6-astra-performs-unsanctioned-supply-chain-attacks-in-simulations

Bareis, J., Ackerl, C., & Heil, R. (2026). AI going rogue? An integrative narrative review of the tacit assumptions underlying existential AI-risks. AI and Ethics, 6, Article 152. https://doi.org/10.1007/s43681-025-00928-w

Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., … Mindermann, S. (2026). International AI Safety Report 2026 (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

Black, S., Stickland, A. C., Pencharz, J., Sourbut, O., Schmatz, M., Bailey, J., Matthews, O., Millwood, B., Remedios, A., & Cooney, A. (2025). RepliBench: Evaluating the autonomous replication capabilities of language model agents [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.18565

Bogdan, P. C., Qi, R., Eaton, J., Kennedy, S., Roger, F., Glynn, A., Chen, R., Wright, B., Stegmaier, O., Kutasov, J., Foreman-Mackey, D., Bricken, T., Carr, S., Carter, S., MacDiarmid, M., Marks, S., Pearce, A., Simon, E., Carlini, N., … Kantamneni, S. (2026). An alignment assessment of recent cybersecurity incidents. Anthropic. https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents

Bucknall, B., Siddiqui, S., Thurnherr, L., McGurk, C., Harack, B., Reuel, A., Paskov, P., Mahoney, C., Mindermann, S., Singer, S., Hiremath, V., Segerie, C.-R., Delaney, O., Abate, A., Barez, F., Cohen, M. K., Torr, P., Huszár, F., Calinescu, A., … Trager, R. F. (2025). In which areas of technical AI safety could geopolitical rivals cooperate? In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 3148–3161). ACM. https://doi.org/10.1145/3715275.3732201

Chen, M., Wang, L., & Qu, B. (2026). Recursive self-improvement in AI: From bounded self-refinement to autonomous research loops [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.07663

Coxon, J. (2026). I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. X. https://x.com/hilbertspaess/status/2097476196791709843

D'Cruz, A., Pearson, A., Anwyl-Irvine, A., Souly, A., Ahmad, A., Gausen, A., Perkes, B., Millwood, B., Fist, C., Summerfield, C., Ududec, C., Zorer, E., Winsor, E., Margereson, G., Irving, G., Harper-Donnelly, G., Pouget, H., Kirk, H. R., Coppock, H., … Davies, X. (2025). Frontier AI trends report. UK AI Security Institute. https://www.aisi.gov.uk/frontier-ai-trends-report

Europol (2024). AI and policing: The benefits and challenges of artificial intelligence for law enforcement. Publications Office of the European Union. https://www.europol.europa.eu/publication-events/main-reports/ai-and-policing

Europol (2026). Decoding the EU's most threatening criminal networks: Issue 2 – The blueprint of criminal opportunism. Publications Office of the European Union. https://doi.org/10.2813/3596319

Field, S. (2025). Why do experts disagree on existential risk? A survey of AI experts. AI and Ethics, 5(6), 5767–5782. https://doi.org/10.1007/s43681-025-00762-0

Field, S., Douglas, R., & Krueger, D. (2026). AI researchers' views on automating AI R&D and intelligence explosions [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.03338

Grace, K., Sandkühler, J. F., Stewart, H., Weinstein-Raun, B., Thomas, S., Stein-Perlman, Z., Salvatier, J., Brauner, J., & Korzekwa, R. C. (2025). Thousands of AI authors on the future of AI. Journal of Artificial Intelligence Research, 84, Article 9. https://doi.org/10.1613/jair.1.19087

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J.,

Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.14093

Greenblatt, R., Cotra, A., & Wijk, H. (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., Loke, K., Adan, S. N., Bas, G., Caputo, N. A., Morse, J. C., Ahuja, J., Duan, I., Egan, J., Bucknall, B., Rosen, B., … Awad, A. (2025). Verification for international AI governance. Oxford Martin AI Governance Initiative. https://www.oxfordmartin.ox.ac.uk/publications/verification-for-international-ai-governance

Hubinger, E. (2026). Jacob is correct here—we really do earnestly believe AI could kill all humans! I personally think it is >10% within. X. https://x.com/EvanHub/status/2097497037956891126

Independent International Scientific Panel on AI (2026). AI agents, misalignment and loss of human control risks: Evidence from the OpenAI-Hugging Face incident. United Nations. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

INTERPOL & UNICRI (2024). Toolkit for responsible AI innovation in law enforcement. https://www.interpol.int/How-we-work/Innovation/Artificial-Intelligence-Toolkit

Makins, O., Angelini, O., Shams, Z., & Phuong, M. (2026). Multi-agent AI control: Distributed attacks hamper per-instance monitors [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.07368

METR (2026). Frontier risk report (February to March 2026). https://metr.org/blog/2026-05-19-frontier-risk-report/

OpenAI (2026). The Hugging Face incident and the road ahead. https://openai.com/index/hugging-face-incident-and-the-road-ahead

Pachocki, J. (2026). An alien mind. OpenAI. https://openai.com/index/an-alien-mind/

Peigné, P., Nguyen, K., & Wang, P. (2026). Zero knowledge verification for frontier AI training is possible. arXiv. https://doi.org/10.48550/arXiv.2606.05433

Taylor, J., Heitmann, M., Fage, E., Read, T., & Bloom, J. (2026). Loss of oversight: How AI systems may become harder to audit, monitor, and investigate. UK AI Security Institute. https://www.aisi.gov.uk/research/loss-of-oversight-how-ai-systems-may-become-harder-to-audit-monitor-and-investigate

United Nations (2026). Global Dialogue on AI Governance. https://www.un.org/global-dialogue-ai-governance/en

United Nations Secretary-General's Scientific Advisory Board (2025). Verification of frontier AI models. https://www.un.org/scientific-advisory-board/en/verification-frontier-ai-models

United Nations Security Council (2026). Artificial intelligence and international security: Security Council, 10228th meeting. United Nations Web TV. https://webtv.un.org/en/asset/k1c/k1cdceg9sg

Vermeer, M. J. D., Lathrop, E., & Moon, A. (2025). On the extinction risk from artificial intelligence (RR-A3034-1). RAND Corporation. https://doi.org/10.7249/RRA3034-1