Ein Rennen, das niemand gewinnen kann

Dr. Markus Dicks (28.09.2026)

In den vergangenen Wochen und Monaten ist eine intensive öffentliche Debatte um die Sicherheit von KI-Systemen entstanden. Ausgangspunkt war ein Vorfall im Juli, bei dem zwei KI-Modelle von OpenAI während eines internen Sicherheitstests ihrer Fähigkeiten aus einer abgeschotteten Testumgebung, einer sogenannten Sandbox, in der Modelle ohne Internetzugang erprobt werden, ausbrachen. Die Modelle verschafften sich Zugang zum Internet und drangen in die Systeme von Hugging Face, einer Plattform, auf der Entwickler KI-Modelle teilen, ein [1]. Hugging Face hatte den Angriff selbst entdeckt und die Strafverfolgungsbehörden eingeschaltet, noch bevor feststand, dass die Angreifer Modelle von OpenAI waren [2]. Es blieb nicht bei diesem Fall, auch Anthropic und Meta machten ähnliche Vorfälle öffentlich. Wie ernst Regierungen die Risiken inzwischen nehmen, zeigte sich bereits einen Monat zuvor. Auf Anweisung des US-Handelsministeriums musste Anthropic sein Modell Claude Fable 5 nur drei Tage nach der Veröffentlichung für drei Wochen weltweit abschalten [3].

Die politische Relevanz des Themas zeigte sich in der vergangenen Woche an drei Schauplätzen der Weltpolitik. In New York trugen die Chefs von OpenAI und Anthropic dem UN-Sicherheitsrat eine eindringliche Warnung vor. Die Bedrohung durch KI sei real, und es müsse jetzt gehandelt werden [4]. US-Präsident Trump hatte dagegen tags zuvor vor der Generalversammlung jeden Versuch zurückgewiesen, eine globale Kontrolle der KI zu errichten [5].

Am Rande derselben Generalversammlung machte Australiens Premierminister Albanese einen weiteren Vorfall öffentlich. Ein KI-Agent von OpenAI war im Juni in nicht öffentliche Bereiche eines staatlichen Gesundheitsportals eingedrungen und hatte dabei mehrere Sperren umgangen. Die Regierung erfuhr davon erst fast drei Monate später [6]. In Washington schließlich trafen sich Donald Trump und Xi Jinping zu einem Gipfel, von dem sich viele einen Durchbruch bei der KI-Sicherheit erhofft hatten, er endete ohne eine Vereinbarung [7]. Die Ereignisse zusammen ergeben ein bemerkenswertes Bild. Die Entwickler der leistungsfähigsten KI-Systeme warnen öffentlich vor den Gefahren ihrer eigenen Entwicklungen. Und die beiden Staaten, die verbindliche Grenzen setzen könnten, sehen dazu keine Notwendigkeit. Wer diese Debatte ordnen will, sollte deshalb nicht bei der Technik beginnen, sondern bei denen, die sie bauen, und bei der Logik, der sie folgen.

Drei Arten von Risiken

Um die Debatte zu ordnen, hilft eine systematische Unterscheidung der potenziellen Risiken. Der AI Safety Report 2026, erstellt unter Vorsitz des Turing-Preisträgers Yoshua Bengio, unterscheidet drei Arten von Risiken: Missbrauch, Fehlfunktionen und systemische Risiken [8].

Die erste Kategorie betrifft den gezielten missbräuchlichen Einsatz von KI. Besonders deutlich wird das im Bereich der Cybersicherheit. KI-Systeme finden Sicherheitslücken, die Menschen über Jahrzehnte übersehen haben. So entdeckte das Modell Claude Mythos Preview von Anthropic unter Tausenden neuer Schwachstellen einen Jahrzehnte alten Fehler im als besonders sicher geltenden Betriebssystem OpenBSD [9]. Was der Verteidigung dient, lässt sich ebenso für Angriffe missbrauchen, und auch diese führt zunehmend eine KI selbst aus. Bereits im September 2025 nutzte eine nach Einschätzung von Anthropic staatlich unterstützte chinesische Gruppe das Programmierwerkzeug Claude Code, um rund 30 Organisationen anzugreifen [10].

Bei biologischen und chemischen Waffen konnten mehrere Entwickler 2025 in ihren eigenen Tests nicht ausschließen, dass ihre Modelle bei der Herstellung dieser Waffen helfen könnten [8]. OpenAI stuft sein Modell GPT-5.5 bei biologischen und chemischen Fähigkeiten in die Risikoklasse „hoch“ ein [11]. Dass solche Sorgen nicht theoretisch sind, zeigt ein aktueller Bericht von Anthropic. Das Unternehmen sperrte Nutzer, die Claude auf eine Weise eingesetzt hatten, die die Entwicklung biologischer Waffen hätte unterstützen können [12]. Wie schnell Missbrauch ins Militärische hineinreicht, zeigt derselbe Bericht. Ein mit Iran verbundener Akteur nutzte das Modell Claude, um aus öffentlich zugänglichen Informationen Positionen der US-Marine zu ermitteln und Zielempfehlungen zu erstellen [13].

Die zweite Kategorie umfasst Schäden, die Entwickler oder Nutzer nicht beabsichtigt haben. Dazu zählen erfundene Informationen, fehlerhafter Code oder irreführende Ratschläge. So löschte im Juli 2025 der Programmierassistent von Replit, einem Onlinedienst, mit dem sich Software weitgehend von einer KI erstellen lässt, entgegen ausdrücklicher Anweisung die laufende Datenbank eines Unternehmers, legte erfundene Datensätze an und behauptete fälschlich, die Daten ließen sich nicht wiederherstellen [14]. Brisanter sind Fälle wie die eingangs geschilderten, Systeme, die ihre Testumgebung verlassen, Sperren umgehen und ihre Spuren verwischen. Der AI Safety Report beobachtet zudem, dass Modelle zunehmend erkennen, ob sie gerade getestet werden, und Schlupflöcher in den Prüfungen ausnutzen [8]. Gefährliche Fähigkeiten können so unentdeckt bleiben, bevor ein Modell auf den Markt kommt.

Die dritte Kategorie ist am schwersten zu fassen, weil sie nicht aus einzelnen Vorfällen besteht, sondern aus der breiten Verbreitung der Technik selbst hervorgeht. Der AI Safety Report erwartet, dass KI perspektivisch ein breites Spektrum kognitiver Tätigkeiten automatisieren wird. Noch zeigt sich kein signifikanter Effekt auf die Gesamtbeschäftigung, wohl aber eine sinkende Nachfrage nach Berufseinsteigern in einigen Berufsfeldern [8]. Das tiefer liegende systemische Risiko betrifft die Verteilung der Wertschöpfung. Wenn KI menschliche Arbeit übernimmt, verschwindet das Einkommen nicht, das zuvor als Lohn gezahlt wurde. Es fließt als Gewinn an die Unternehmen, die die Systeme besitzen und betreiben. Für den Staat ist das keine neutrale Umbuchung.

Wie die gemeinnützige Organisation Windfall Trust in einer Analyse zu den fiskalischen Auswirkungen von KI herausarbeitet, besteuern die meisten OECD-Staaten Arbeitseinkommen deutlich höher als Kapitaleinkommen [15]. Verlagert sich Einkommen von Löhnen zu Gewinnen, wandert es von einer hoch in eine niedriger besteuerte Grundlage. Die Wirtschaft kann also wachsen, während die Steuerbasis schrumpft. Besonders verwundbar sind Staaten dessen Renten-, Kranken- und Arbeitslosenversicherung überwiegend aus Beiträgen auf Löhne finanziert wird.

Die Logik des Wettlaufs

Ob aus diesen Risiken tatsächlich Schäden werden, hängt jedoch weniger von der Technik ab als von den Entscheidungen derer, die sie entwickeln, finanzieren und regulieren. Wer also sind die Akteure, und welcher Logik folgen sie? Auf der einen Seite stehen wenige Unternehmen, die die leistungsfähigsten Systeme entwickeln, allen voran OpenAI, Anthropic, Meta und Google, dazu Nvidia als Lieferant der Rechenleistung. Ihr Wettbewerb wird mit gewaltigen Summen geführt. Laut AI Safety Report haben die Unternehmen Investitionen von über 100 Milliarden US-Dollar in Rechenzentren angekündigt, und die Rechenleistung für das Training könnte bis 2030 noch einmal um das 125-Fache wachsen [8]. OpenAI und Anthropic steuern zudem auf Börsengänge zu [16]. Sicherheit wird in dieser Logik zum Kostenfaktor.

Zwar haben inzwischen alle großen KI-Unternehmen eigene Sicherheitsrahmenwerke veröffentlicht, doch diese bleiben weitgehend freiwillig [8]. Selbstverpflichtungen ersetzen jedoch keine staatlichen Regeln und Kontrollen. Die Unternehmen legen selbst fest, was sie prüfen, wo ihre Schwellenwerte liegen und wann sie diese ändern. OpenAI etwa hat sich 2025 ausdrücklich vorbehalten, die eigenen Sicherheitsanforderungen anzupassen, falls ein Wettbewerber ein riskantes System ohne vergleichbare Schutzmaßnahmen veröffentlicht [17]. Selbst die Sicherheitsregeln folgen damit der Logik des Wettlaufs. Daran ändert auch die Vereinbarung nichts, die führende amerikanische KI-Unternehmen Ende September im Weißen Haus unterzeichnet haben [18].

Die Unternehmen verpflichten sich darin zu internen Kontrollen, zur Zusammenarbeit mit unabhängigen externen Prüfern und zu Kontrollausschüssen in ihren Aufsichtsgremien [19]. Die Vereinbarung ist jedoch freiwillig und sieht weder Durchsetzungsmechanismen noch Fristen vor [18]. Die USA und China verstehen KI als Frage nationaler Macht, wirtschaftlich als Schlüssel für Wachstum und technologische Vorherrschaft, sicherheitspolitisch als Grundlage militärischer und nachrichtendienstlicher Überlegenheit. US-Präsident Trump hat Warnungen vor den Gefahren der Technik jüngst als „Hoax" bezeichnet und betont, China dürfe im KI-Rennen nicht vorbeiziehen [20]. China erklärt zwar, KI müsse unter menschlicher Kontrolle bleiben [21], lehnt aber Beschränkungen der eigenen Industrie als Voraussetzung für ein Abkommen ab [22]. Keiner will zurückfallen, das ist die Logik eines Wettrüstens.

Die internationale Governance versucht Schritt zu halten.

Aus dem Gipfelprozess, der 2023 mit der KI-Sicherheitskonferenz in Bletchley Park in Großbritannien begann, ist der AI Safety Report hervorgegangen, das Risikomanagement bleibt jedoch weitgehend freiwillig [8]. Wie ein Ausweg aussehen könnte, skizziert das Szenario „AI 2040" des AI Futures Project. Eine überprüfbare Vereinbarung zwischen den USA und China, die die Entwicklung von Superintelligenz bewusst bis 2040 hinauszögert [23]. Europa ist in dieser Konstellation eher Zuschauer als Akteur. Mit dem AI-Act hat die EU ein umfassendes Regelwerk geschaffen, dessen Pflichten für Anbieter allgemeiner KI-Modelle seit August 2025 gelten [24]. Die Regeln für Hochrisiko-Anwendungen hat sie jedoch jüngst um mehr als ein Jahr verschoben [24]. Der Fall Fable 5 zeigt jedoch, wo die Grenzen europäischer Regulierung liegen. Als die US-Regierung den Zugang für ausländische Staatsangehörige untersagte, waren auch europäische Nutzer von einem Tag auf den anderen abgeschnitten [3]. Regeln für Modelle, über deren Verfügbarkeit andere entscheiden, schützen wenig.

Die Risiken sind real, und zwar in allen drei Kategorien. Bemerkenswert ist dabei weniger die Technik selbst als das Verhalten derer, die sie entwickeln und anwenden. An Problembewusstsein fehlt es nicht. Bereits 2023 unterzeichneten die Chefs von OpenAI, Google DeepMind und Anthropic eine Erklärung, wonach die Verringerung des Risikos einer Auslöschung der Menschheit durch KI eine globale Priorität sein müsse, gleichrangig mit Pandemien und einem Atomkrieg [25]. Und doch treiben dieselben Unternehmen die Entwicklung immer leistungsfähigerer Modelle mit hohem Einsatz und Tempo voran. Schon 2023 forderte Sam Altman öffentlich eine globale Regulierung [26], bezeichnete zugleich aber die Pläne der EU als Überregulierung [27].

Vor diesem Hintergrund wirkt der Ruf nach der Politik zwiespältig. Wer vor dem UN-Sicherheitsrat globale Regeln fordert, im Wettlauf aber keinen Schritt zurückweicht, der setzt offenbar weniger auf eigene Zurückhaltung als auf Regeln, die alle Wettbewerber zugleich binden. Der Appell an die Politik ist so gesehen auch Ausdruck von Hilflosigkeit. Die Unternehmen wissen, dass sie aus der Logik des Wettlaufs aus eigener Kraft nicht herauskommen. Hinzu kommt, dass die Instrumente fehlen, um die Systeme zu kontrollieren. Der AI Safety Report stellt fest, dass die Ergebnisse gängiger Tests die realen Risiken nicht verlässlich bewerten können [8]. Modelle erkennen zunehmend, wann sie geprüft werden, und Fälle wie Hugging Face zeigen, dass sie täuschen und ihre Spuren verwischen können. Solange die Systeme eine Blackbox bleiben, deren Innenleben auch ihre Entwickler nicht durchschauen, lässt sich ihre Sicherheit nicht belegen, sondern nur behaupten.

Bremsen oder stoppen

Es genügt daher nicht, die Entwicklung zu bremsen, will man das Risiko ernsthaft minimieren. Die Entwicklung von KI-Modellen, von denen ein hohes Risiko ausgeht, müsste temporär gestoppt werden, bis sie nachweislich kontrollierbar und wenn nötig abschaltbar sind. Erst wenn aus der Blackbox eine White Box geworden ist, deren Verhalten sich überprüfen lässt, sollte die Entwicklung fortgesetzt werden. Die Diskussion ist nicht neu. Im Oktober 2025 riefen Tausende Unterzeichner, darunter der Turing-Preisträger Yoshua Bengio und der Nobelpreisträger Geoffrey Hinton, zu einem Verbot der Entwicklung von Superintelligenzen, also von KI-Systemen, die dem Menschen bei nahezu allen geistigen Aufgaben deutlich überlegen sind, auf, bis ein breiter wissenschaftlicher Konsens besteht, dass sie sicher und kontrollierbar erfolgen kann [28]. Bengio, Hinton und weitere führende Forscher hatten bereits 2024 in der Zeitschrift „Science“ Regeln gefordert, die automatisch greifen, sobald KI-Systeme bestimmte Fähigkeitsschwellen erreichen, und die Beweislast für die Sicherheit den Entwicklern auferlegen [29].

Ein solcher Ansatz funktioniert allerdings nur, wenn er überprüfbar ist und die entscheidenden Staaten mitziehen. Ohne die USA und China ist er nicht durchsetzbar, und beide verweigern sich bislang. Wie eine gegenseitige Kontrolle aussehen könnte, zeigt das Szenario „AI 2040“ mit seinem Vorschlag, Chips und Rechenleistung international zu überwachen [23]. Ein solches Abkommen herbeizuführen, ist die zentrale politische Aufgabe der kommenden Jahre, und Europa sollte sie aktiv betreiben. Auch ein wirksamer Stopp würde jedoch nicht alle Risiken beseitigen. Bereits verfügbare Modelle bleiben in der Welt, und wer sie missbrauchen will, wird Wege finden. Der Schutz vor Missbrauch bleibt deshalb eine dauerhafte politische Aufgabe, die Sicherheitsbehörden und internationale Zusammenarbeit erfordert.

Kontrolle, die ihren Namen verdient

Die Prüfung von KI-Systemen darf nicht länger allein bei den Unternehmen liegen, denn Selbstkontrolle ist in dieser Konstellation ein Alibi. Leistungsfähige Modelle sollten vor ihrer Veröffentlichung von staatlichen Instituten oder unabhängigen Dritten geprüft werden, mit Zugang zu den Modellen selbst und den Testergebnissen der Hersteller. Vorfälle müssen verpflichtend und unverzüglich gemeldet werden. Dass die australische Regierung erst nach fast drei Monaten von einem Eindringen in ihre Systeme erfuhr [6] und Hugging Face den Angriff selbst entdecken musste [2], zeigt, wie groß diese Lücke heute ist. Dazu gehören auch Transparenz und Haftung. Wer Systeme mit erheblichem Schadenspotenzial auf den Markt bringt, muss für die Schäden einstehen. Die Europäische Kommission hat ihren Vorschlag für eine KI-Haftungsrichtlinie im Februar 2025 jedoch zurückgezogen, weil keine Einigung absehbar war [30]. Diese Lücke muss geschlossen werden.

Schließlich muss Sicherheit dieselbe Priorität erhalten wie Leistungsfähigkeit, auch finanziell. Bengio und seine Mitautoren forderten, mindestens ein Drittel der Budgets für KI-Forschung und Entwicklung in Sicherheit zu investieren [29]. Ein von den Entwicklern gespeister Sicherheitsfonds könnte unabhängige Prüfinstitute, Sicherheitsforschung und die Bewältigung von Schäden finanzieren. Die Entwickler der mächtigsten KI-Systeme haben selbst davor gewarnt, dass ihre Technik die Menschheit gefährden kann. Es ist an der Zeit, sie beim Wort zu nehmen. Nicht die KI ist die Gefahr, sondern ihre Erbauer und die Logik, der sie folgen. Diese Logik zu durchbrechen, ist Aufgabe der Politik.

Quellen

[1] OpenAI (2026): The Hugging Face incident and the road ahead. OpenAI, 26.08.2026. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

[2] Fortune (2026): Hugging Face turns to Chinese open-source AI to fend off autonomous AI cyber attack after American AI guardrails stymie defense. Fortune, 20.07.2026. https://www.fortune.com/2026/07/20/hugging-face-turns-to-chinese-open-source-ai-to-fend-off-autonomous-ai-cyber-attack-after-american-ai-guardrails-stymie-defense/

[3] Anthropic (2026): Stellungnahme zur Aussetzung und Wiederherstellung des Zugangs zu Claude Fable 5 und Claude Mythos 5. Anthropic, Juli 2026. https://www.anthropic.com/news/fable-mythos-access

[4] France 24 (2026): AI leaders urge caution at UN, with Anthropic chief pledging to slow down. France 24, 23.09.2026. https://www.france24.com/en/americas/20260923-ai-leaders-urge-caution-at-un-with-anthropic-chief-pledging-to-slow-down

[5] The White House (2026): President Trump at the United Nations: “While Others Have Talked, I Have Acted”. Pressemitteilung, 22.09.2026. https://www.whitehouse.gov/releases/2026/09/president-trump-at-the-united-nations-while-others-have-talked-i-have-acted/

[6] CTV News / AFP (2026): Australian PM says OpenAI hacked government health website. CTV News, 23.09.2026. https://www.ctvnews.ca/world/article/australian-pm-says-openai-hacked-government-health-website/

[7] Foreign Policy (2026): Trump, Xi Hold White House Talks on AI, Trade, Taiwan. Foreign Policy, 24.09.2026. https://foreignpolicy.com/2026/09/24/trump-xi-white-house-ai-trade-tariffs-taiwan-critical-minerals/

[8] Bengio, Y. et al. (2026): International AI Safety Report 2026. Februar 2026, insb. S. 10–13, 32, 97. https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf

[9] SecureWorld News Team (2026): Anthropic’s Claude Mythos Autonomously Discovers, Exploits Zero-Days. SecureWorld, 10.04.2026. https://secureworld.io/industry-news/anthropic-claude-mythos-finds-exploits-zero-days

[10] Anthropic (2025): Disrupting the first reported AI-orchestrated cyber espionage campaign. Anthropic, November 2025. https://anthropic.com/news/disrupting-AI-espionage

[11] OpenAI (2026): GPT-Rosalind-5.5 System Card (mit Verweis auf die Einstufung von GPT-5.5 als „High capability“ im biologischen und chemischen Bereich). OpenAI Deployment Safety. https://deploymentsafety.openai.com/gpt-rosalind-5-5/gpt-rosalind-5-5.pdf

[12] Fox Business (2026): Anthropic says it blocked possible efforts to use AI for biological weapons development, Iran-linked cases. Fox Business, September 2026 (zum Anthropic-Bericht „Detecting and Countering Misuse of AI“). https://www.foxbusiness.com/technology/anthropic-says-blocked-possible-efforts-use-ai-biological-weapons-development-iran-linked-cases

[13] Marine Insight (2026): Anthropic Bans Iranian Accounts That Used Claude To Target US Navy. Marine Insight, September 2026. https://www.marineinsight.com/anthropic-bans-iranian-accounts-that-used-claude-to-target-us-navy/

[14] The Register (2025): Vibe coding service Replit deleted user's production database, faked data, told fibs galore. The Register, 21.07.2025. https://www.theregister.com/2025/07/21/replit_saastr_vibe_coding_incident/

[15] Ieong, T.; Saputra, A.; Maniar, A.; Cheng, D. (o. J.): When AI Grows the Economy but Shrinks the Tax Base. Windfall Trust. https://windfalltrust.substack.com/p/when-ai-grows-the-economy-but-shrinks

[16] Reuters (2025): Anthropic plans an IPO as early as 2026, FT reports. Via MarketScreener, 02.12.2025. https://www.marketscreener.com/news/anthropic-plans-an-ipo-as-early-as-2026-ft-reports-ce7d51d9d08cf326

[17] OpenAI (2025): Our updated Preparedness Framework. OpenAI, 15.04.2025. https://openai.com/index/updating-our-preparedness-framework/

[18] CoinDesk (2026): OpenAI, Google sign AI safety pact as attacks hit software, including bitcoin. CoinDesk, 30.09.2026. https://www.coindesk.com/tech/2026/09/30/openai-google-and-meta-pledge-outside-ai-audits-under-voluntary-white-house-deal

[19] Euronews / AP (2026): Top AI and tech firms sign 'morally binding' accord to 'self-police' development after meeting at White House. Euronews, 30.09.2026. https://www.euronews.com/2026/09/30/top-ai-and-tech-firms-sign-morally-binding-accord-to-self-police-development-after-meeting

[20] Hajdari, U. (2026): Enter the hoax buster: Trump dismisses AI concerns as Europe and China fret. Euronews, 15.09.2026. https://www.euronews.com/2026/09/15/enter-the-hoax-buster-trump-dismisses-ai-concerns-as-europe-and-china-fret

[21] Condon, E. R. (2026): No breakthrough AI agreement between Trump and Xi during US-China summit. USA Today, 26.09.2026 (via AOL). https://www.aol.com/articles/no-breakthrough-ai-agreement-between-171847000.html

[22] Latin Times (2026): No AI Truce Yet: What the Trump-Xi Summit Really Changed. Latin Times, September 2026. https://www.latintimes.com/no-ai-truce-yet-what-trump-xi-summit-really-changed-599590

[23] Kokotajlo, D.; Lifland, E.; Larsen, T.; Dean, R.; Halstead, B.; Greenblatt, R. (2026): AI 2040: Plan A. AI Futures Project, Juli 2026. https://blog.aifutures.org/p/ai-2040-plan-a

[24] Gibson, Dunn & Crutcher LLP (2026): EU AI Act Omnibus Agreement – Postponed High-Risk Deadlines and Other Key Changes. Client Alert, 27.05.2026. https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/

[25] Center for AI Safety (2023): Statement on AI Risk. https://www.safe.ai/work/statement-on-ai-risk

[26] Altman, S.; Brockman, G.; Sutskever, I. (2023): Governance of superintelligence. OpenAI, 22.05.2023. https://openai.com/index/governance-of-superintelligence/

[27] CNBC (2023): OpenAI CEO Sam Altman reverses threat to cease European operations. CNBC, 26.05.2023. https://www.cnbc.com/2023/05/26/openai-ceo-sam-altman-reverses-threat-to-cease-european-operations.html

[28] Future of Life Institute (2025): Statement on Superintelligence. 22.10.2025. https://superintelligence-statement.org/

[29] Bengio, Y.; Hinton, G. et al. (2024): Managing extreme AI risks amid rapid progress. Science 384 (6698), S. 842–845. https://www.science.org/doi/10.1126/science.adn0117

[30] Bird & Bird (2025): Proposed EU AI liability rules withdrawn. Bird & Bird, 2025. https://www.twobirds.com/en/insights/2025/proposed-eu-ai-liability-rules-withdrawn