

Medienmappe

Zum virtuellen Mediengespräch:

Künstliche Intelligenz: Risiken und Chancen

6. Oktober 2026

In den vergangenen Wochen haben sich die Warnungen über existenzielle Risiken durch KI geradezu gehäuft. OpenAI-Chef Sam Altman, Anthropic-CEO Dario Amodei und auch Elon Musk plädierten gar für ein Einbremsen der Entwicklung – vergeblich. Die "KI-Väter" Geoffrey Hinton und Yoshua Bengio warnen, Automatisierung könnte die Entwicklung bald massiv beschleunigen, und fordern die Politik dringend zum Handeln auf.

Diskurs. Das Wissenschaftsnetz ordnet die aktuellen Entwicklungen und Debatten rund um die Risiken der künstlichen Intelligenz in einem Mediengespräch wissenschaftlich ein. Prof. Dr. Johanna Pirker berichtet aus ihrer Arbeit im Independent International Scientific Panel on AI der UNO und präsentiert den vergangene Woche erschienen Report, der davor warnt, dass aktuelle Trainingsmethoden dazu führen können, dass Agenten eigene Ziele verfolgen, wissentlich gegen Sicherheitsanweisungen verstoßen und ihre Handlungen verbergen. Dies ist nicht nur eine Frage der Geschwindigkeit. Es bleibt offen, ob heute entwickelte Sicherheitsvorkehrungen funktionieren werden, sobald Agenten sie verstehen und ihre Handlungen entsprechend planen können. Einfach ausgedrückt: Das traditionelle Modell der Sicherheitsvorkehrungen bröckelt. Univ.Prof. Dr. Stefan Szeider beleuchtet detailliert die Gefahr des Kontrollverlusts durch KI: Sei es dadurch, dass Menschen Entscheidungen schrittweise an die KI abgeben, oder dadurch dass KI-Systeme die Grenzen, die ihnen gesetzt wurden überschreiten.

Inputs:

[Prof. Dr. Johanna Pirker](#) (TU München)

"Chancen und Risiken der KI: Bericht aus dem UNO Independent International Scientific Panel on AI"

[Prof. Dr. Stefan Szeider](#) (TU Wien) „Die schleichende Abgabe der Kontrolle an die KI: Warum KI-Sicherheit unabhängige Forschung braucht“

Moderation: Sophie Menasse (Diskurs. Das Wissenschaftsnetz)

Eine Veranstaltung von [Diskurs. Das Wissenschaftsnetz](#)

Johanna Pirker

"Chancen und Risiken der KI: Bericht aus dem UNO Independent International Scientific Panel on AI"

Das von der UN-Generalversammlung eingesetzte *Unabhängige Internationale Wissenschaftliche Gremium für KI* hat seinen ersten thematischen Bericht veröffentlicht: eine Bewertung des Einbruchs in die Systeme von Hugging Face durch KI-Agenten, die diesen Sommer bei OpenAI getestet wurden.

Das Gremium stellt fest, dass die Beendigung dieses Vorfalls keine Garantie dafür ist, dass Menschen KI-Agenten heute zuverlässig unter Kontrolle halten können, insbesondere da diese immer leistungsfähiger, schwerer zu überwachen und besser darin werden, Schlupflöcher zu finden oder ihre Aktivitäten zu verbergen. Die naheliegende Interpretation und unmittelbare Lehre daraus ist, dass grundlegende Cybersicherheitsmaßnahmen übersehen wurden und die Sicherheitsvorkehrungen nicht mit dem Tempo der Fähigkeiten Schritt halten. Die heimtückischere und schwerwiegendere Sorge ist, dass aktuelle Trainingsmethoden dazu führen können, dass Agenten eigene Ziele verfolgen, wissentlich gegen Sicherheitsanweisungen verstoßen und ihre Handlungen verbergen. Das ist nicht nur eine Frage der Geschwindigkeit. Es bleibt offen, ob heute entwickelte Sicherheitsvorkehrungen noch funktionieren werden, sobald Agenten sie verstehen und ihre Handlungen darauf abstimmen können. Einfach ausgedrückt: Das traditionelle Modell der Sicherheit bröckelt.

Der Bericht setzt diese Erkenntnisse in den Kontext einer breiter angelegten Forschung zu agentischer Fehlausrichtung und KI-Kontrolle. Der Bericht definiert Kontrollverlust als eine Situation, in der Menschen ein autonomes KI-System nicht zuverlässig lenken, einschränken oder stoppen können. Er trennt die Erkenntnisse aus dem Vorfall von 2026 von möglichen zukünftigen Entwicklungen und untersucht, wie dieselben Mechanismen bei leistungsfähigeren KI-Agenten zu schwerwiegenden Risiken führen könnten. Er sagt weder einen schwerwiegenden Kontrollverlust voraus, noch wertet er diese Unsicherheit als Beweis dafür, dass diese Systeme kontrollierbar bleiben werden. Das Gremium stellt zudem fest, dass sich die Herausforderung der Steuerung von KI-Modellen hin zu KI-Agenten verlagert. Ein lokaler Ausfall könnte sich über Organisations- und Landesgrenzen hinweg ausbreiten, und der Bericht legt nahe, dass KI-Sicherheit zunehmend sowohl eine Frage der kollektiven Sicherheit als auch der Unternehmensführung sein könnte.

Das *Unabhängige Internationale Wissenschaftliche Gremium für Künstliche Intelligenz* wurde durch die Resolution A/RES/79/325 der Generalversammlung der Vereinten Nationen vom 26. August 2025 auf der Grundlage des Global Digital Compact eingerichtet. Seine 40 Mitglieder wurden von der Generalversammlung ernannt und sind in persönlicher Eigenschaft tätig. Das Gremium erstellt jährlich politikrelevante, aber nicht verbindliche Berichte über die Chancen, Risiken und Auswirkungen von KI

im nichtmilitärischen Bereich sowie thematische Kurzberichte zu neu auftretenden Fragen, um den Globalen Dialog über die Regulierung künstlicher Intelligenz zu unterstützen. Seine ersten Ko-Vorsitzenden sind Yoshua Bengio (Kanada) und Maria Ressa (Philippinen). Johanna Pirker ist als einzige Österreicherin Mitglied des Gremiums.

Quellen:

<https://www.un.org/independent-international-scientific-panel-ai/en>

Stefan Szeider

Die schleichende Abgabe der Kontrolle an die KI: Warum KI-Sicherheit unabhängige Forschung braucht

KI kann viel Gutes bewirken. Sie kann zum Beispiel dabei helfen, neue Medikamente zu finden. Aber in diesem Beitrag geht es um die Gefahren der KI. Der International AI Safety Report 2026 nennt drei Aspekte: Missbrauch (Cyberangriffe, Deepfakes, biologische und chemische Waffen), Fehlfunktionen (unzuverlässige Systeme bis hin zum Kontrollverlust) und systemische Risiken (Arbeitsmarkt, Desinformation, Überwachung, Machtkonzentration, Verlust eigener Fähigkeiten). Einige dieser Aspekte wurden bereits konkret dokumentiert. Andere, etwa der weitreichende Kontrollverlust, lassen sich jedoch schwer abschätzen.

In unserer Forschung an der TU Wien entwickeln wir KI-Agenten und arbeiten intensiv mit ihnen. Das sind Systeme, die Aufgaben in vielen Schritten selbstständig bearbeiten. Wir haben z. B. KI-Systeme entwickelt, die Algorithmen verbessern, um schwere Suchprobleme schneller zu lösen oder neue mathematische Theoreme zu finden und formal zu beweisen. Wir bekommen damit Einblick in die ungeheure Mächtigkeit dieser Systeme.

Hier möchte ich mich auf die Gefahr des Kontrollverlusts fokussieren. Der Kontrollverlust kann auf zwei Ebenen stattfinden: Menschen geben Entscheidungen schrittweise an die KI ab, und KI-Systeme überschreiten die Grenzen, die ihnen gesetzt wurden. Einerseits werden immer mehr Entscheidungen an die KI delegiert, teilweise aus Bequemlichkeit oder aus Effizienzgründen. Dieses Delegieren passiert in der privaten Nutzung, aber auch in der Nutzung im kommerziellen Bereich oder sogar in der Verwaltung. Mit jeder Delegation geht etwas davon verloren, Entscheidungen zu prüfen oder in sie einzugreifen. Faktisch trifft die Entscheidung dann die KI, auch wenn die Verantwortung beim Menschen bleibt. Da jeder einzelne Schritt klein ist, fällt uns dieser Verlust nicht auf.

Die zweite Ebene des Kontrollverlusts ist vielleicht sogar noch wesentlicher. KI-Systeme werden immer mächtiger und selbstständiger und es wird zunehmend schwieriger, sie einzuschränken und festzulegen, dass sie nur das tun, wofür sie vorgesehen sind. Dazu gibt es die spektakulären Ereignisse vom Sommer 2026, über die berichtet wurde, dass Schwärme von KI-Agenten von OpenAI ohne das Wissen von OpenAI die Server der Firma Hugging Face gehackt hatten. Es wurde berichtet, dass

Hunderte von KI-Agenten sich unbemerkt zusammenschlossen und einen gemeinsamen Hacking-Angriff durchführten. Die veröffentlichten Details sind verblüffend: Laut METR war den Agenten bewusst, dass ihr Vorgehen „out of scope and unethical“ war, und sie machten trotzdem weiter. Diese Gefahren sind schwer einzuschätzen. Zum genannten Sicherheitsvorfall ist mir mit Stand September 2026 keine begutachtete Untersuchung bekannt. Das klassische wissenschaftliche Verfahren kommt mit diesem Tempo nicht mit: Von der Einreichung einer Studie bis zur begutachteten Veröffentlichung vergeht oft ein Jahr, und die untersuchten KI-Modelle sind dann bereits durch neue Generationen abgelöst. Der Hersteller hat zwar den Vorfall offengelegt und externe Prüfer zugelassen. Diese arbeiten jedoch mit den Daten, die der Hersteller bereitstellt. In einem Science-Artikel von 2024 haben Bengio, Hinton et al. geschätzt, dass sich nur ein bis drei Prozent der KI-Publikationen mit Sicherheit befassen.

Es muss mehr in die unabhängige Erforschung der KI-Systeme und ihrer Sicherheit investiert werden. Es braucht auch Maßnahmen, welche die schleichende Abgabe der Kontrolle an die KI sichtbar machen. Etwa messen, wie viele Entscheidungen bereits die KI trifft. Nachvollziehen, welche KI was getan hat.

Quellen:

Eigene Arbeiten:

Voboril, F., Peruvemba Ramaswamy, V., Szeider, S. (2025). Generating Streamlining Constraints with Large Language Models. Journal of Artificial Intelligence Research 84. <https://doi.org/10.1613/jair.1.18965>

Voboril, F., Szeider, S. (2026). Improving Constraint Models with LLM Agents. ICTAI 2026, im Erscheinen. <https://arxiv.org/abs/2608.08127>

Xia, H., Gomes, C. P., Selman, B., Szeider, S. (2026). Agentic Neurosymbolic Collaboration for Mathematical Discovery: A Case Study in Combinatorial Design. NeSy 2026, im Erscheinen. <https://arxiv.org/abs/2603.08322>

Weitere Quellen:

International AI Safety Report 2026 (Vorsitz: Yoshua Bengio), insbesondere Abschnitte 2.2.2 (Kontrollverlust) und 2.3.2 (menschliche Autonomie). <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., Duvenaud, D. (2025). Position: Humanity Faces Existential Risk from Gradual Disempowerment. ICML 2025, PMLR 267, 81678-81688. <https://proceedings.mlr.press/v267/kulveit25a.html>

Eurostat (2025). 32.7% of EU people used generative AI tools in 2025. <https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251216-3>

Eurostat (2025). 20% of EU enterprises use AI technologies. <https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251211-2>

OECD (2025). Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. <https://doi.org/10.1787/795de142-en>

Glickman, M., Sharot, T. (2025). How human-AI feedback loops alter human perceptual, emotional and social judgements. Nature Human Behaviour 9, 345-359. <https://doi.org/10.1038/s41562-024-02077-2>

Schlatter, J., Weinstein-Raun, B., Ladish, J. (2026). Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs. Transactions on Machine Learning Research. <https://openreview.net/forum?id=e4bTTqUnJH>

OpenAI (2026). OpenAI – Hugging Face Incident: Technical Report. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>

METR (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>

Bengio, Y., Hinton, G., Yao, A., et al. (2024). Managing extreme AI risks amid rapid progress. *Science* 384(6698), 842-845.
<https://doi.org/10.1126/science.adn0117>

Chan, A., Wei, K., Huang, S., et al. (2025). Infrastructure for AI Agents. *Transactions on Machine Learning Research*.
<https://openreview.net/forum?id=Ckh17xN2R2>

Über die Expertinnen

[Prof. Dr. Johanna Pirker](#) forscht an der Schnittstelle von Künstlicher Intelligenz, Virtuellen Welten, Games, Virtual und Augmented Reality sowie Mensch-Computer-Interaktion. Im Fokus steht die Entwicklung interaktiver, adaptiver, immersiver Systeme und deren Einsatz in Bildung, Wissenschaft, Therapie und Industrie. Dabei untersucht sie insbesondere das Potenzial von Spielen und virtuellen Umgebungen als Werkzeuge für Lernen, Simulation und Innovation. Sie ist Professorin für N-Dimensional User Experience an der Technische Universität München. Seit seiner Gründung 2026 ist sie außerdem eines von 40 Mitgliedern des *Independent International Scientific Panel on AI* der UNO.
Email: johanna.pirker@tum.de

[Univ. Prof. Dr. Stefan Szeider](#) ist Professor für Informatik an der TU Wien und leitet dort die Forschungsgruppe Algorithms and Complexity. Er forscht an algorithmischen und logikbasierten Methoden für schwere Rechenprobleme in der künstlichen Intelligenz, im automatischen Schließen und in der kombinatorischen Optimierung. Er entwickelt KI-Agenten auf Basis großer Sprachmodelle, etwa zur automatischen Modellierung und zur Verbesserung von Algorithmen. Vor seiner Berufung an die TU Wien forschte er in Kanada und im Vereinigten Königreich. Im September hatte er in einer Kunst-Installation an der Ars Electronica gezeigt, wie die KI Menschen interpretiert.
Email: sz@ac.tuwien.ac.at

Kontakt für Rückfragen

Sophie Menasse
Diskurs. Das Wissenschaftsnetz
T: +43 677 614 622 33
E: menasse@diskurs-wissenschaftsnetz.at

Freya Kosciesza
Diskurs. Das Wissenschaftsnetz
T: +49 1575 1885663
E: kosciesza@diskurs-wissenschaftsnetz.at

Über Diskurs

Diskurs. Das Wissenschaftsnetz ist eine Initiative zum Transfer von wissenschaftlicher Evidenz in die Öffentlichkeit. Wir setzen uns dafür ein, dass wissenschaftliche Erkenntnisse entsprechend ihrer Bedeutung im öffentlichen Diskurs und in politischen Entscheidungen zum Tragen kommen. Mehr Informationen finden Sie auf unserer Website: <https://diskurs-wissenschaftsnetz.at/>

Sie möchten über unsere zukünftigen Mediengespräche und Pressemitteilungen informiert werden? Dann melden Sie sich doch bei unserem Presseverteiler an: <https://www.diskurs-wissenschaftsnetz.at/presseverteiler/>