## Quellen & Weiterführende Lektüre


[1] Shapira, N., Riedl, C., Bau, D. et al. (38 Autorinnen und Autoren) (2026, Februar). *Agents of Chaos*. arXiv Preprint, arXiv:2602.20021. Northeastern University, Harvard University, MIT, Stanford University, Carnegie Mellon University u.a. URL: https://arxiv.org/abs/2602.20021

[2] Shapira, N., Riedl, C., Bau, D. et al. (2026). *Agents of Chaos – Interactive Report*. Baulab, Northeastern University. Interaktiver Forschungsbericht mit Fallstudien, Protokollen und Timelines. URL: https://agentsofchaos.baulab.info/report.html

[3] Shapira, N. / Northeastern University (Redaktion: Cesareo, A.) (2026, 9. März). *They wanted to put AI to the test. They created agents of chaos*. Northeastern University News. Pressebericht zur Studie. URL: https://news.northeastern.edu/2026/03/09/autonomous-ai-agents-of-chaos/

[4] Bengio, Y., Clare, S., Prunkl, C., Murray, M. et al. (über 100 Expertinnen und Experten) (2026, 3. Februar). *International AI Safety Report 2026*. Veröffentlicht im Auftrag von 29 Nationen, der EU, OECD und den Vereinten Nationen. Geleitet von Prof. Yoshua Bengio (Université de Montréal / Mila). Vollständiger Bericht. URL: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[5] Bengio, Y. et al. (2026). *International AI Safety Report 2026 – Executive Summary*. Dreiseitige Zusammenfassung der Kernbefunde zu AI-Fähigkeiten, Risiken und Sicherheitsmassnahmen. URL: https://internationalaisafetyreport.org/publication/2026-report-executive-summary

[6] RAND Europe (2025). *Examining Risks and Response for AI Loss of Control Incidents*. RAND Europe Research Report. Cambridge, UK. URL: https://www.rand.org/randeurope/research/projects/2025/examining-risks-and-response-for-ai-loss-of-control-incidents-cm.html

[7] Casper, S. et al. (MIT und weitere Institutionen) (2026, Februar). *The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems*. arXiv Preprint, arXiv:2602.17753. URL: https://arxiv.org/html/2602.17753v1

[8] Aembit (Hrsg.) (2026). *Agentic AI Guardrails: What They Are and How to Implement Them*. Fachbeitrag. Quelle für Gartner-Prognose (40 % Enterprise-Agenten bis Ende 2026) und McKinsey-Erhebung. URL: https://aembit.io/blog/agentic-ai-guardrails-for-safe-scaling/

[9] OWASP (Open Worldwide Application Security Project) (2025). *LLM01:2025 Prompt Injection*. OWASP Top 10 for LLM Applications & Generative AI, Version 2025. Internationales Sicherheits-Rahmenwerk. URL: https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[10] Underhill, K. / eSecurity Planet (2025, Dezember). *AI Agent Attacks in Q4 2025 Signal New Risks for 2026*. eSecurity Planet. Analyse der Lakera-Q4-2025-Daten zu AI-Agenten-Angriffen. URL: https://www.esecurityplanet.com/artificial-intelligence/ai-agent-attacks-in-q4-2025-signal-new-risks-for-2026/

[11] OpenAI (2025). *Detecting and Reducing Scheming in AI Models*. OpenAI Technical Blog. Forschungsbericht zu Sandbagging und Anti-Scheming-Training in Frontier-Modellen. URL: https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/