

Handreichung zum Deep Dive
Cyber-Bedrohungen in Zeiten von KI

09. September 2026 | Werder (Havel)

Prof. Dr. Volker Wittpahl
Klaipeda University (KU), Litauen
Mail: volker.wittpahl@ku.lt
Tel.: +49 179 14 333 28

TEIL 1: Kurze Übersicht zur KI-Entwicklung und KI-Anwendungen

KI-Einsatz in der Wirtschaft

- Übersicht zu KI-Einsatz in der Wirtschaft
 - Quelle: Sequoia, <https://sequoiacap.com/article/generative-ai-act-two/>

TEIL 2: Cyber-Bedrohungen in Zeiten von KI

KI-Cybersecurity

KI-induzierte Cybersicherheitsrisiken

- Prompt Injection in Review-Prozessen von wissenschaftlichen Veröffentlichungen
 - Quelle: Heise vom 07.07.2025, „Winzige Schriftart, weiße Farbe: Versteckte KI-Prompts in Fachartikeln entdeckt“, <https://www.heise.de/news/Gib-nur-positive-Bewertungen-Prompts-fuer-KI-Tools-in-Fachartikeln-versteckt-10476429.html>
- Zuverlässigkeitskrise von Agentic AI
 - Quelle: Tiwari, U.; Whittaker, M. (2025): AI Agent, AI Spy, December 2025, 39C3, Hamburg, <https://media.ccc.de/v/39c3-ai-agent-ai-spy#t=148>
- Auswirkungen auf die Cybersicherheit von Organisationen durch die Entwicklung im Bereich Künstlicher Intelligenz
 - Quelle: Bundesamt für Sicherheit in der Informationstechnik (BSI), BITS-B Nr. 2026-262788-1032, Version 1.0, 22.06.2026
https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2026/2026-262788-1032_bits.html

Die unkontrollierbare Autonomie von Agentic AI

- 2024: ChatGPT o1
 - Quelle: BR24, 10.12.2025, <https://www.br.de/nachrichten/netzwelt/chatgpt-rebelliert-luegt-und-klont-sich-selbst,UWXd8KC>
- 2025: Anthropic
 - Quelle: Handelsblatt, 22.05.2025, <https://www.handelsblatt.com/technik/ki/kuenstliche-intelligenz-ki-software-greift-in-test-zu-erpressung-aus-selbstschutz/100130310.html>
- 2026: Agentic AI attackiert Software-Entwickler
 - Quelle: Heise Online, 14.02.2026, <https://www.heise.de/news/WTF-KI-Agent-attackiert-Entwickler-oeffentlich-nach-abgelehnter-Code-Aenderung-11176583.html>
- 2026: Open-AI's KI-Modell hackt Hugging Face
 - Quelle: Heise Online, 23.07.2026, <https://www.heise.de/news/Autonome-KI-hackt-Hugging-Face-Weckruf-fuer-IT-Sicherheit-und-Politik-11374152.html>
 - Quelle: OpenAI, OpenAI – Hugging Face Incident, Technical Report, <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
 - Quelle: METR Redwood Research, 26.08.2026, Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
 - Quelle: Dwarkesh Patel, 30.08.2026, The Rise and Fall of Agent Civilizations, <https://www.dwarkesh.com/p/openai-huggingface>

KI-Cybersafety

Wachsende Trends auf der individuellen Ebene

- In den kommenden Monaten:
 - Verlust des Vertrauens in digitale Nachrichten und digitale Quellen
 - Verlust des Vertrauens in die digitale Kommunikation
 - Verlust des Vertrauens in die Privatsphäre

Vertrauensverlust in die digitale Kommunikation und ihre Quellen

- Betrüger gibt sich in Telefonat mit Giffey als Vitali Klitschko aus
 - Quelle: FAZ vom 25.06.2022, <https://www.faz.net/aktuell/politik/ausland/mann-gibt-sich-in-telefonat-mit-giffey-als-klitschko-aus-18127999.html>
- Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'
 - Quelle: CNN vom 04.02.2024, <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>

KI-Tag der Wirtschaft

- Beispiel: Komplette „automatisierte“ Phishing-Operationen
 - Quelle: Candid Wüest, „GenAI and Cybercrime: Separating Fact from Fiction“, November 2024, DeepSec 2024 Wien
- AI Slop
 - Quelle: c't 2026/05, Seite 55 | Bild: Martina Bruns / heise medien

Vertrauensverlust der Privatsphäre und Bedrohung des freien Willens

Sie sind der begehrteste Trainingsdatensatz

- Corporate Surveillance in Everyday Life
 - Quelle: Cracked Labs, 2017, <https://crackedlabs.org/en/corporate-surveillance>
- Die Akte Xandr - Ein tiefer Blick in den Abgrund der Datenindustrie
 - Quelle: 36c3, 2023, <https://media.ccc.de/v/37c3-11974-die-akte-xandr-ein-tiefer-blick-in-den-abgrund-der-datenindustrie>
- Databroker Files: Wie uns Apps und Datenhändler der Massenüberwachung ausliefern
 - Quelle: 38c3, 2024, <https://media.ccc.de/v/38c3-databroker-files-wie-uns-apps-und-datenhndler-der-massenberwachung-ausliefern>

Psychische und kognitive Risiken durch KI

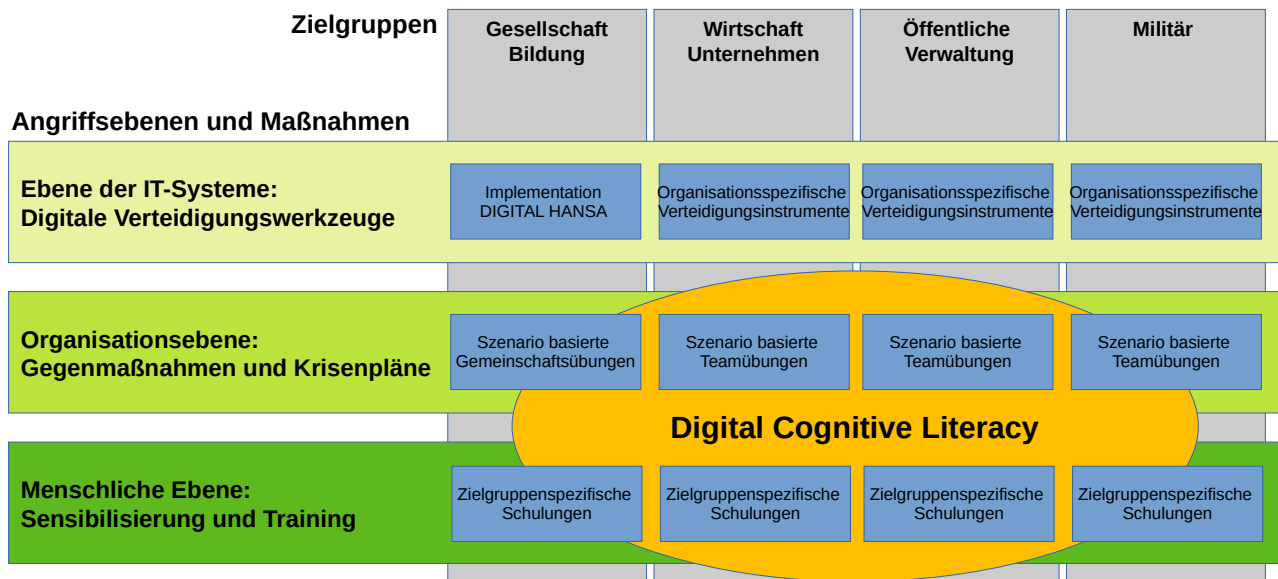
- Durch KI verursachte kognitive Belastung und organisatorische Auswirkungen
 - Quelle: Westover (2026): „When AI Assistance Becomes Cognitive Overload“ 10.70175/hclreview.2020.31.4.5
- Schlaue KI – schrumpfendes Hirn
 - Quelle: c't Heft 10, 2025, Seite 124 - 127
- Wie manipulativ ist KI? So beeinflussen Sprachmodelle unsere Überzeugungen
 - Quelle: c't Heft 3, 2025, Seite 133 - 137
- 2025: Virtuelle Influencer: Die Top 10 in 2025
 - Quelle: Storyclash vom 12.11.2024, <https://www.storyclash.com/blog/de/virtuelle-influencer/>
- KI-Einfluss auf moralische Entscheidungen und Verantwortung
 - Salatino, A., Prével, A., Caspar, E. et al. (2025): Influence of AI behavior on human moral decisions, agency, and responsibility. Sci Rep 15, 12329 (2025). <https://doi.org/10.1038/s41598-025-95587-6>
- Emotionale Bindung von Nutzern zu KI-Systemen
 - Kloiber, M. (2026): Einnehmend: Wie KI-Chatbots emotionale Bindungen erzeugen, Int. Bastian Schiller, 31. Januar 2026, <https://www.deutschlandfunk.de/einnehmend-wie-ki-chatbots-emotionale-bindungen-erzeugen-int-bastian-schiller-100.html>
- Belegte gesundheitliche Beeinträchtigungen
 - Molavi-Vasse'i (2026): Companion AI – Studie zu Risiken von KI-Begleitsystemen und ihre regulatorische Erfassung. Zentrum für Digitalrechte und Demokratie, <https://digitalrechte.de/publikationen/companion-ai>
 - Augustin (2025): „KI-assoziierte Psychose: Erkenntnisse aus ersten Fällen“. Der Nervenarzt 96, Nr. 7 (2025): 699–701.

Der Einsatz von KI in der Cyber- und Informationskriegsführung

- Russland manipuliert wohl westliche Chatbots für seine Propaganda
 - Quelle: Spiegel vom 17.03.2025, <https://www.spiegel.de/ausland/chatgpt-und-perplexity-ai-russland-manipuliert-westliche-chatbots-fuer-seine-propaganda-a-7e276236-cac3-4f35-8ad4-40eaba1c8caf>
- Bedrohung der Demokratie
 - Schroeder, Daniel Thilo et al. (2026): How malicious AI swarms can threaten democracy, Science 391, 354 (2026), DOI: 10.1126/science.adz1697
- Policy Paper „Cognitive Security“
 - Bıçakcı, Salih (2025): Cognitive Security in the age of AI: Building national resilience against synthetic influence, Policy Paper No. 4, Jean Monnet Centre of Excellence, 2025, <https://resaid.bilgi.org.tr/wp-content/uploads/2025/11/Cognitive-Security-Salih-Bicakci.pdf> (siehe Anhang zum Handout)

TEIL 3: Cybersecurity wird Cognitive Security

Nationale Cognitive Security Strategieskizze



Organisationsspezifische Verteidigungsinstrumente

- Agentic AI Framework OpenClaw
 - Quelle: <https://openclaw.ai/>
- Leitfaden zur Implementierung eines „Lokalen KI-Systems“
 - Wittpahl, V. Steinebach, M. Endres, C. Holling, J. (2026): Schnell pilotieren, sicher skalieren - Lokale KI im Unternehmen und weitere Alternativen, BSI Allianz für Cybersicherheit (März 2026), https://www.allianz-fuer-cybersicherheit.de/SharedDocs/Downloads/Webs/ACS/DE/Kreise/Expertenkreis_KI_Pilotierung.pdf?__blob=publicationFile&v=2

Implementation des DIGITAL HANSA Ansatzes: Ein EU Trusted Internet zur Verhältnisprävention

- DIGITALE HANSE
 - Quelle: Wittpahl, V., DIGITAL HANSA - Digital Gated Communities lead into a Trusted EU Internet, White Paper Draft 1.0 (März 2026)