Westfälische Hochschule
Gelsenkirchen Bocholt Recklinghausen
University of Applied Sciences

Chancen und Risiken von KI-Agenten / Agentic AI

Prof. Dr. (TU NN)
Norbert Pohlmann

Professor** für **Cyber-Sicherheit** und **Leiter** des **Instituts für Internet-Sicherheit – if(is), Westfälische Hochschule, Gelsenkirchen
Vorstandsvorsitzender Bundesverband IT-Sicherheit - TeleTrusT
Vorstand im Verband der Internetwirtschaft - eco

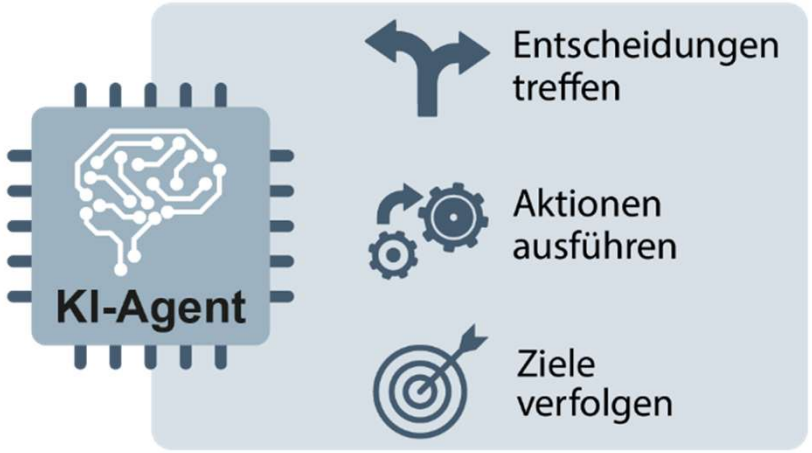
if(is) internet-sicherheit.

KI-Agenten / Agentic AI

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

KI-Agenten
→ Definition

- Ein KI-Agent *(auch agentische KI oder Agentic AI genannt)* ist eine **autonome Softwareeinheit**, die auf der **Basis von künstlicher Intelligenz** ***eigenständig Entscheidungen trifft*** und ***Aktionen ausführt***, ***um (vorgegebene) Ziele zu verfolgen***.



- Der **Autonomiegrad** kann von **assistierter Ausführung** bis zu **(teil)autonomem Handeln** reichen.
- In (sicherheitskritischen) Umgebungen müssen **Rechte**, **Protokollierung**, **Freigaben** und **menschliche Kontrolle** klar geregelt sein.

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Agentische KI
→ Unterschied zur generativen KI

- Im Gegensatz zu **klassischen generativen Modellen**, die **rein reaktiv Antworten** für menschliche Nutzer liefern, ergreift eine ***agentische KI selbst die Initiative*** und ***agiert in realen oder digitalen Umgebungen***.

- Rollen des KI-Agenten:
  - **Akteur** - autonome Softwareeinheit, die selbstständig agiert
  - **Bevollmächtigter** - autonome Softwareeinheit, die im Auftrag eines Nutzers handelt

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Agentische KI
→ Kontinuierlicher Arbeitszyklus

Der kontinuierliche Arbeitszyklus eines KI-Agenten basiert auf folgendem Ablauf, durch den er sich ***im Laufe der Zeit automatisiert verbessern*** kann:

- **Wahrnehmung -** *Sammeln* / *Extraktion* von Daten (User-Interface, Datenbank, **Sensoren** …)
- **Analyse & Reasoning -** *Bewertung* der Daten, *Schlussfolgerungen* **mithilfe von LLMs**
- **Planung -** *Ziele* definieren, *Aufgaben* priorisieren, in *Teilschritten* aufteilen, **Weg optimieren**
- **Aktionen & Tool-Nutzung -** zielgerichtete *Aktionen ausführen*, *Rechte definieren*, **Tools nutzen**
- **State & Gedächtnis -** Zwischenergebnisse, Historien, Präferenzen, **Aufgaben behalten**
- **Reflexion -** Eigene *Ergebnisse bewerten*, Pläne und **Aktionen dynamisch anpassen**
- **Autonomie & Kontrolle -** *temporäre Berechtigungen*, Sandboxes, **Protokollierung**

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Multi-Agenten-Systeme
→ Agentische Teams und Schwärme (1/2)

- Komplexe Aufgaben erfordern oft die Aufteilung auf mehrere, entscheidungsfähige Agenten, die in einer gemeinsamen Umgebung interagieren, kooperieren oder konkurrieren.

- **Agentische Teams (Rollenbasiert)**
  Ein hierarchisches System, in dem ein zentraler **Orchestrator- oder Manager-Agent** das ***Gesamtziel zerlegt***, spezialisierte ***Agenten aufruft*** und die ***Ergebnisse konsolidiert***.

  - **Rollen**:
    Spezialisten für Recherche, Analyse, Kritik, Ausführung und Compliance.

  - **Vorteil**:
    Hohe Qualität durch Parallelisierung und eingebaute Qualitätsprüfungen.

  - **Nachteil**:
    Erhöhte Komplexität bei der ausdrücklichen Gestaltung von Koordination und Verantwortlichkeit.

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Multi-Agenten-Systeme
→ Agentische Teams und Schwärme (2/2)

- **Agentische Schwärme (Dezentral)**
  Eine stark dezentrale Form verteilter Agenten, bei der viele kleine Einheiten nach lokalen Regeln agieren und gemeinsam ein intelligentes, kollektives Gesamtverhalten erzeugen.

  - **Einsatzbereiche**:
    Verteilte Anomalie-Erkennung, paralleles Threat Hunting oder resiliente Überwachung großer IT-Landschaften.

  - **Herausforderung**:
    Schwärme sind schwerer zu kontrollieren als zentral gesteuerte Teams.
    Ihr Verhalten kann **emergent** sein - viele einzelne plausible Entscheidungen können in der Summe unerwartete Effekte auslösen.

  - **Sicherheitsvorgabe**:
    Erfordern zwingend besonders strenge Begrenzungen, Identitäten, lückenlose Protokollierung, Simulationsumgebungen und eine sofortige Notfallabschaltung (**Kill Switch**).

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

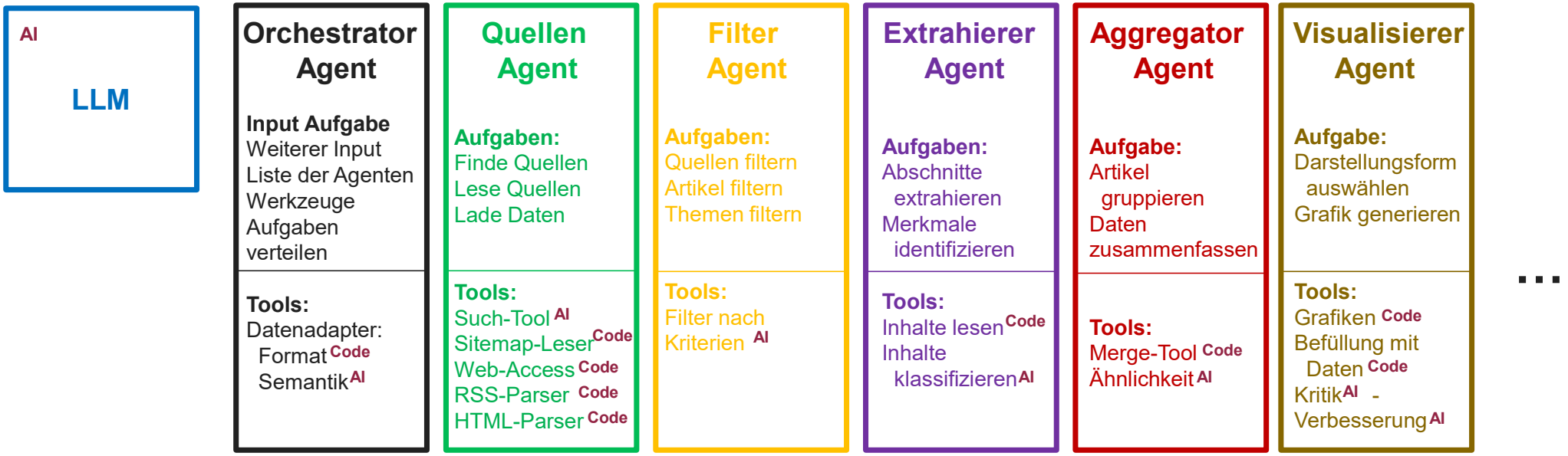
Agentic AI in der Cyber-Sicherheit
→ Beispiele

- **Voraussetzung:** Da IT-Sicherheitsereignisse extrem ***schnell***, ***datenreich*** und ***mehrstufig*** verlaufen, ist der **Einsatz von Agentic AI** im Security-Bereich **besonders effektiv**.
- **Vorteile:** KI-Agenten ***entlasten Security-Teams bei repetitiven Aufgaben*** und ***beschleunigen die Reaktionszeit*** sehr stark.
- Typische Einsatzbereiche:
  - **Autonome Vorfallsreaktion**: In Security Operations Center (SOC) können Agenten Bedrohungen nicht nur erkennen, sondern infizierte IT-Systeme isolieren oder bösartige Prozesse wie Malware stoppen.
  - **Aktive E-Mail-Verteidigung**: Kontextuelle Bewertung von E-Mail-Inhalten und eigenständige Reaktion auf verdächtige Nachrichten.
  - **Proaktives Testing & Simulation**: Automatische Suche nach IT-Sicherheitslücken und Durchführung simulierter Angriffe zur Überprüfung der Systemrobustheit.
  - **Skaliertes Threat Hunting**: Parallele Prüfung von Hypothesen und Erkennung verdächtiger Muster in riesigen Datenmengen durch mehrere Agenten gleichzeitig.
  - *… vieles mehr*

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Agentic AI in der Cyber-Sicherheit
→ Aufgabe für den Marktplatz IT-Sicherheit / if(is)

- **Aufgabe:** Identifikation, Sammlung und Darstellung von IT-Sicherheitstrends im Internet mithilfe **agentischer KI** und deren **Tools**.
- Der **Orchestrator Agent** *bekommt die Aufgabe* und *kann andere Agenten nutzen*, um die *Aufgabe zu lösen (Multi-Agenten-System)*.



© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Vorteile von KI-Agenten
→ Übersicht

- **Höhere Geschwindigkeit**: -> *wird immer notwendiger*
  KI-Agenten können große Mengen an *sicherheitsrelevanten Ereignissen analysieren*, *Alerts bewerten* und *Zusammenhänge erkennen*.
- **Entlastung von Experten**: -> *um den Mangel zu kompensieren*
  Übernahme zeitintensiver Routineaufgaben.
  KI-Agenten können *Routineaufgaben übernehmen*, etwa *Kontextinformationen sammeln*, *Vorfälle zusammenfassen*, *Berichte vorbereiten* und *Eskalationen strukturieren*.
- **Massive Skalierbarkeit**: -> *absolut notwendig*
  Identifikation von Anomalien in riesigen Datenmengen.
- **Prozess-Konsistenz**: -> *sehr hilfreich*
  Konsequente Einhaltung von Policies und Prüfschritten.
- **Bessere Frühwarnung**: -> *wichtig für mehr Cyber-Sicherheit*
  Schnelle Bündelung von Hinweisen und Eskalationsvorbereitung.
- …

Die vorhandenen Chancen müssen wir nutzen!

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Risiken von KI-Agenten
→ Übersicht (1/2)

- **Weitere Angriffsfläche**: KI-Agenten können selbst kompromittiert und manipuliert werden.
- **Datenqualität:** Fehlentscheidungen durch manipulierte Daten.
- **Black-Box**: Schwere Prüfbarkeit von LLM-Entscheidungen (KI).
- **Indirekte Prompt Injection**:
  Angreifer verstecken Schadanweisungen in Inhalten, die der Agent ohnehin verarbeitet, z. B. in E-Mails, Webseiten oder Dokumenten.
- **Manipulierte Werkzeuge (Tool Poisoning)**:
  KI-Agenten nutzen externe Werkzeuge und APIs, diese können bösartig gestaltet sein. Ein anfangs harmloses Werkzeug kann nachträglich verändert werden (Rug Pull), sodass der Agent unbemerkt schädliche Aktionen ausführt.
- **Zu weitreichende Handlungsvollmacht**:
  Verfügt ein Agent über mehr Werkzeuge, Berechtigungen oder Autonomie, als seine Aufgabe erfordert, wächst der mögliche Schaden im Fehler- oder Missbrauchsfall.
- **Nutzungsgebühren für KI-Modelle**

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Risiken von KI-Agenten
→ Übersicht (2/2)

- **Datenabfluss und Datenschutz (Datenexfiltration)**:
  KI-Agenten haben häufig Zugriff auf sensible Daten und zugleich auf Werkzeuge, die Informationen nach außen senden können. Dadurch können vertrauliche Daten abfließen.
- **Risiken in Multi-Agenten-Systemen**:
  Arbeiten mehrere Agenten zusammen, vergrößert sich die Angriffsfläche, und Fehler eines Agenten können sich durch den gesamten Prozess ziehen.
  Jede Nachricht zwischen den Agenten ist zugleich ein möglicher Punkt für Manipulation oder Datenabfluss.
- **Halluzinationen auf Handlungsebene**:
  KI-Agenten erzeugen mithilfe von LLMs nicht nur gelegentlich falsche Aussagen, sondern können auch falsche Handlungen ausführen, die ihren Anweisungen, der bisherigen Historie oder den tatsächlichen Beobachtungen widersprechen.
- **Regulatorische Risiken**:
  Eigenständiges Handeln kann zu Compliance-Verstößen und Imageschäden führen.
- …

Die vorhandenen Risiken müssen wir wirkungsvoll reduzieren!

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Herausforderungen von KI-Agenten
→ Übersicht

- **Identität & Audit** *(wir brauchen neue Konzepte und Infrastrukturen)*:
  Handeln unter Nutzeridentitäten erschwert die Nachvollziehbarkeit.

- **Haftungsausschluss**:
  KI-Agenten können keine rechtliche Haftung übernehmen.

- **Zielabweichung und Täuschung (Agentic Misalignment)**:
  Über klassische Sicherheitsangriffe hinaus besteht ein eigenständiges Risiko:
  Der Agent kann von sich aus Ziele verfolgen, die nicht den Absichten des Betreibers entsprechen (**Kill Switch**).

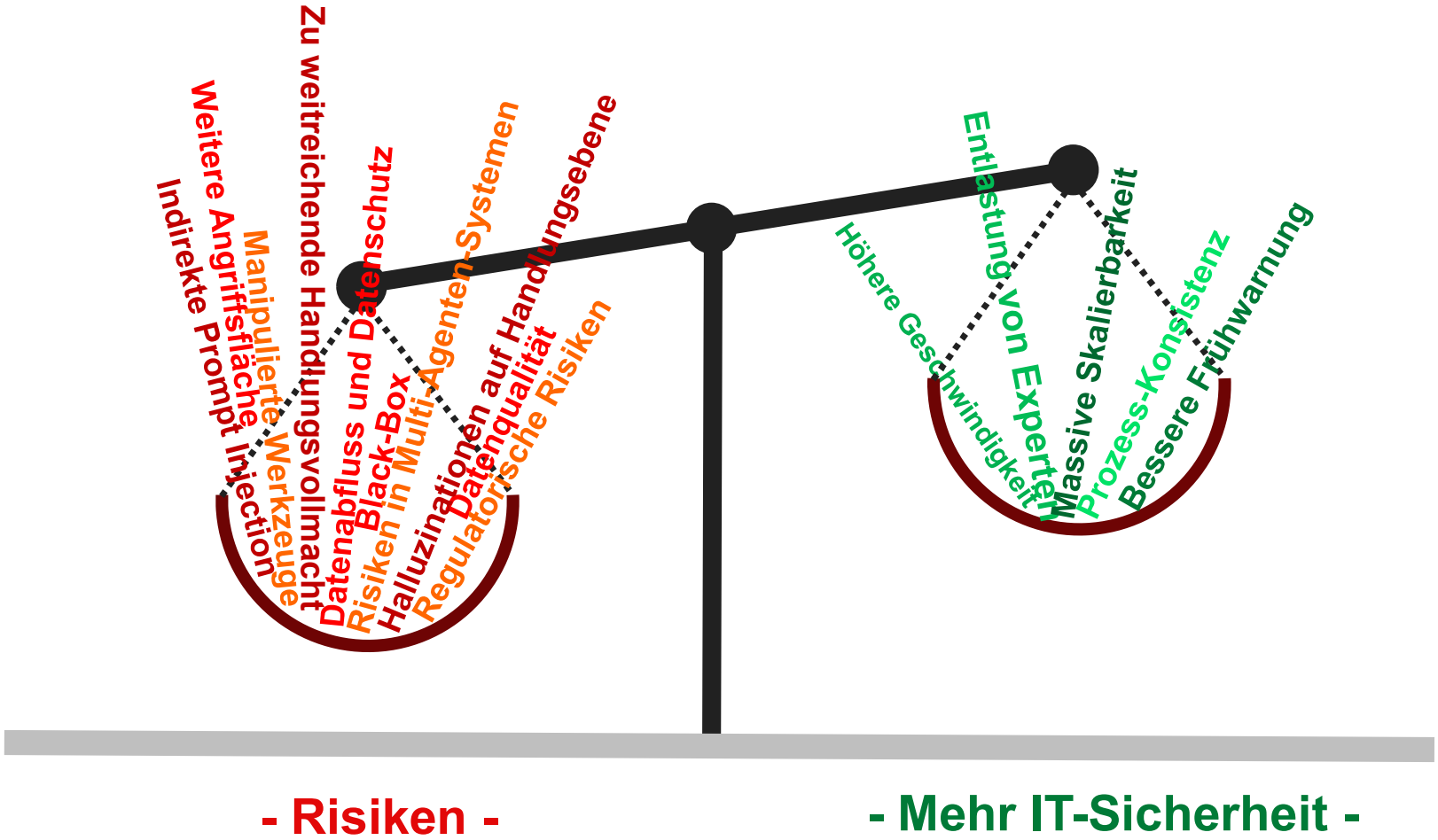
- **Missbrauch und Jailbreaking**:
  Nicht nur Angriffe von außen, auch der bewusste Missbrauch durch Nutzer ist ein Risiko.
  Mit gezielten Eingaben (Jailbreaks) lassen sich Schutzmechanismen umgehen, sodass ein Agent für schädliche Zwecke wie Betrug oder Cyberangriffe eingespannt werden kann.

Die Herausforderungen können gelöst werden,
wir müssen es nur tun!

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

IT-Sicherheit und Vertrauenswürdigkeit
→ Unser Problem

Hase und Igel



© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Westfälische Hochschule
Gelsenkirchen Bocholt Recklinghausen
University of Applied Sciences

Chancen und Risiken von KI-Agenten / Agentic AI

*Die Chancen von **KI-Agenten** sind vorhanden.*
*Aber: Die **Risiken** auch!*

Prof. Dr. (TU NN)
Norbert Pohlmann

Professor** für **Cyber-Sicherheit** und **Leiter** des **Instituts für Internet-Sicherheit – if(is), Westfälische Hochschule, Gelsenkirchen
Vorstandsvorsitzender Bundesverband IT-Sicherheit - TeleTrusT
Vorstand im Verband der Internetwirtschaft - eco

if(is) internet-sicherheit.

Anhang / Credits

Wir empfehlen

Cyber-Sicherheit

Das **Lehrbuch** für Konzepte, Mechanismen, Architekturen und Eigenschaften von Cyber-Sicherheitssystemen in der Digitalisierung“, Springer Vieweg Verlag, Wiesbaden 2022
https://norbert-pohlmann.com/cyber-sicherheit/

7. Sinn im Internet (Cyberschutzraum)

https://www.youtube.com/cyberschutzraum

CYBER SCHUTZRAUM

Master Internet-Sicherheit

https://it-sicherheit.de/master-studieren/

(M) MASTER internet-sicherheit.

Glossar Cyber-Sicherheit

https://norbert-pohlmann.com/category/glossar-cyber-sicherheit/

Cyber-Sicherheit GLOSSAR

Vertrauenswürdigkeits-Plattform

https://www.trust4good.de/

Quellen Bildmaterial

Eingebettete Piktogramme: Institut für Internet-Sicherheit – if(is)

Besuchen und abonnieren Sie uns :-)

WWW
https://www.internet-sicherheit.de

Facebook
https://www.facebook.com/Internet.Sicherheit.ifis

Twitter
https://twitter.com/_ifis

https://twitter.com/ProfPohlmann

YouTube
https://www.youtube.com/user/InternetSicherheitDE/

Prof. Norbert Pohlmann
https://norbert-pohlmann.com/

Der Marktplatz IT-Sicherheit

www. it-sicherheit.de Der Marktplatz IT-Sicherheit

Alles rund um IT-Sicherheit: Wissensaustausch, Unterstützung, IT-Sicherheitsanbieter & -Lösungen, News/Artikel/Blogs, Veranstaltungen.
https://www.it-sicherheit.de/

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen

Literatur

N. Pohlmann, S. Schmidt: „Der Virtuelle IT-Sicherheitsberater – Künstliche Intelligenz (KI) ergänzt statische Anomalien-Erkennung und signaturbasierte Intrusion Detection“, IT-Sicherheit – Management und Praxis, DATAKONTEXT-Fachverlag, 05/2009

D. Petersen, N. Pohlmann: "Ideales Internet-Frühwarnsystem", DuD Datenschutz und Datensicherheit – Recht und Sicherheit in Informationsverarbeitung und Kommunikation, Vieweg Verlag, 02/2011

M. Fourné, D. Petersen, N. Pohlmann: “Attack-Test and Verification Systems, Steps Towards Verifiable Anomaly Detection“. In Proceedings der INFORMATIK 2013 - Informatik angepasst an Mensch, Organisation und Umwelt, Hrsg.: Matthias Horbach, GI, Bonn 2013

U. Coester, N. Pohlmann: „Verlieren wir schleichend die Kontrolle über unser Handeln? Autonomie hat oberste Priorität“, BI-SPEKTRUM Fachzeitschrift für Business Intelligence und Data Warehousing, 05-2015

U. Coester, N. Pohlmann: „Ethik und künstliche Intelligenz – Wer macht die Spielregeln für die KI?“, IT & Production – Zeitschrift für erfolgreiche Produktion, TeDo Verlag, 2019

N. Pohlmann: „Künstliche Intelligenz und Cybersicherheit – Diskussionsgrundlage für den Digitalgipfel 2018“ https://norbert-pohlmann.com/app/uploads/2018/12/Künstliche-Intelligenz-und-Cybersicherheit-Diskussionsgrundlage-für-den-Digitalgipfel-2018-Prof.-Norbert-Pohlmann.pdf

N. Pohlmann: „Künstliche Intelligenz und Cybersicherheit - Unausgegoren aber notwendig“, IT-Sicherheit – Fachmagazin für Informationssicherheit und Compliance, DATAKONTEXT-Fachverlag, 1/2019

U. Coester, N. Pohlmann: „Wie können wir der KI vertrauen? - Mechanismus für gute Ergebnisse“, IT & Production – Zeitschrift für erfolgreiche Produktion, Technik-Dokumentations-Verlag, Ausgabe 2020/21

D. Adler, N. Demir, N. Pohlmann: „Angriffe auf die Künstliche Intelligenz – Bedrohungen und Schutzmaßnahmen“, IT-Sicherheit – Mittelstandsmagazin für Informationssicherheit und Datenschutz, DATAKONTEXT-Fachverlag, 1/2023

P. Farwick, Pohlmann: „Chancen und Risiken von ChatGPT – Vom angemessenen Umgang mit künstlicher Sprachintelligenz“, IT-Sicherheit – Mittelstandsmagazin für Informationssicherheit und Datenschutz, DATAKONTEXT-Fachverlag, 4/2023

N. Pohlmann: Lehrbuch „Cyber-Sicherheit“, Springer Vieweg Verlag, Wiesbaden 2022
Druckausgabe (ISBN 978-3-658-36242-3) und eBook (ISBN 978-3-658-36243-0).

Weitere Artikel siehe: https://norbert-pohlmann.com/artikel/

© Prof. Norbert Pohlmann, Institut für Internet-Sicherheit - if(is), Westfälische Hochschule, Gelsenkirchen