Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Spezielle Angriffsvektoren /-arten von KI (Bedrohungsmodellierung)

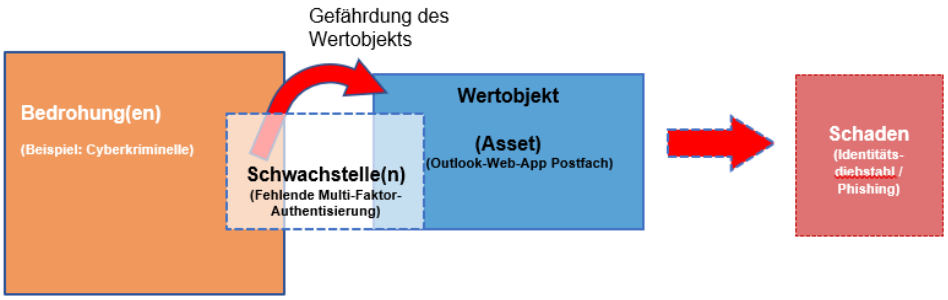
Dr. Rainer Rumpel
AURISCON GmbH
13. März 2025

1

AURISCON BERÄT. PRÜFT. SICHERT.

Bedrohungen von Informationen - Risiko

Risiko ist eine Kombination aus (**B**edrohung, **S**chwachstelle, **A**sset, **S**chaden). R=(B,Ss,A,Sd)



2

Die Norm ISO/IEC 27005:2018 2022 liefert im Anhang A 2.5

AURISCON BERÄT. PRÜFT. SICHERT.

Typische Bedrohungen: ISO/IEC 27005:2022, Anhang A.2.5

- Auszug

Infrastruktur	Bb09	Ausfall der Kühlung
	Bb10	Ausfall des Telekommunikationssystems
	Bb11	Ausfall der Wasserversorgung
Kompromittierung von Information	Bc12	Abfangen von Strahlung
	Bc13	Ausspähen von Information
	Bc14	Abhören von Kommunikation
	Bc15	Diebstahl von Dokumenten
	Bc16	Diebstahl von Geräten
	Bc17	Diebstahl von Datenträgern
	Bc18	Verlust von Dokumenten
	Bc19	Verlust von Geräten
	Bc20	Verlust von Datenträgern
	Bc21	Offenlegung von Information
	Bc22	Manipulation von Information
	Bc23	Datenwiederherstellung bei recycelten oder ausgesonderten Medien
	Bc24	Einspielen von Nachrichten
	Bc25	Informationen oder Produkte aus unzuverlässiger Quelle
	Bc26	Missbrauch personenbezogener Daten
	Bc27	Identitätsdiebstahl
	Bc28	Schadsoftware

3

AURISCON BERÄT. PRÜFT. SICHERT.

Typische Schwachstellen: ISO/IEC 27005:2022, Anhang A.2.5

- Auszug

Hardware	Sc18	Empfindlichkeit bei Spannungsschwankungen
	Sc19	Empfindlichkeit bei Temperaturextremen
	Sc20	Keine Kontrolle des Ereignisprotokolls
	Sc21	Ungeschützter Datenspeicher
	Sc22	Mangelnde Sorgfalt bei der Entsorgung
Software	Sd23	Unzureichendes Testen
	Sd24	Softwarefehler
	Sd25	Sicherheitsmangel
	Sd26	Unangemessene Zugriffsmöglichkeit
	Sd27	Mangelnde Benutzerfreundlichkeit
	Sd28	Dokumentationsmangel
	Sd29	Ungeeignete Parameterkonstellation
	Sd30	Falsches Datum
	Sd31	Unzureichender Authentisierungsmechanismus
	Sd32	Ungeschützte Kennworttabelle
	Sd33	Unnötiger Dienst aktiv
	Sd34	Unreife oder neue Software
	Sd35	Unzureichende Vorgaben für Entwickler

4

AURISCON BERÄT. PRÜFT. SICHERT.

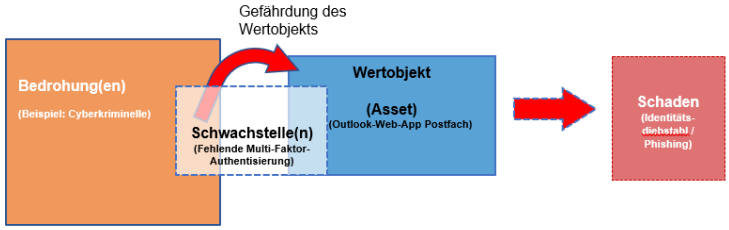
Bedrohungen von Informationen - Risiko

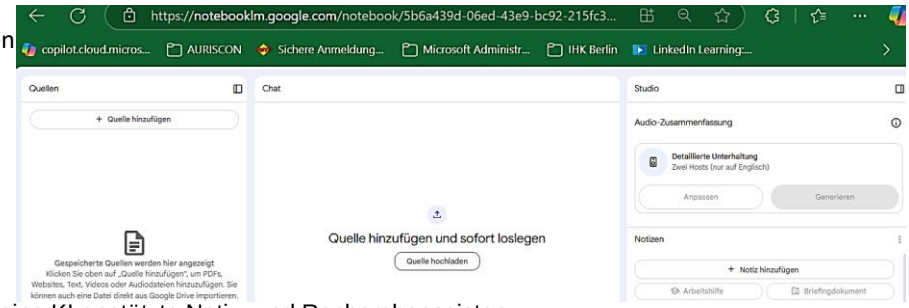
Beispiel für R=(B,Ss,A,Sd)

B: Offenlegung von Information
Ss: Softwarefehler

A: NotebookLM*
Enterprise**

Sd: Abfluss von Geschäftsgeheimnissen





*Google's NotebookLM **ist** eine KI-gestützte Notiz- und Rechercheassistenz.
** NotebookLM Enterprise läuft in einer unternehmensspezifischen Google-Cloud-Umgebung

5

AURISCON BERÄT. PRÜFT. SICHERT.

Erweitertes Risikomodell

Einbeziehung der Informationssicherheits**z**iele

Vertraulichkeit, Integrität, Verfügbarkeit

siehe z.B. *ISO/IEC 27000 Information security management systems — Overview and vocabulary*

Also **R=(B,Ss,A,Sd,Z)**

Beispiel

B: Offenlegung von Information, Ss: Softwarefehler, A: NotebookLM Enterprise, Sd: Abfluss von Geschäftsgeheimnissen

Z: Vertraulichkeit, d.h. Verletzung der Vertraulichkeit

6

AURISCON BERÄT. PRÜFT. SICHERT.

Zweites Beispiel

R=(B,Ss,A,Sd,Z)

B: Abfluss von Information, Ss: Backdoor, A: NotebookLM Enterprise, Sd: Abfluss von Geschäftsgeheimnissen

Das sind „klassische Informationssicherheitsrisiken“.

7

AURISCON BERÄT. PRÜFT. SICHERT.

KI-Spezifische Informationsrisiken

Evasion-Angriffe

- Der Angreifer manipuliert Eingabedaten, sodass das KI-Modell falsche Ausgaben liefert, ohne dass diese Manipulation bei Tests durch den Menschen offensichtlich ist.
- Es werden geringfügige, kaum sichtbare Veränderungen an den Eingabedaten vorgenommen, die das Modell dazu bringt, eine fehlerhafte Klassifizierung vorzunehmen.
- Beispiel: Verkehrsschilder mit eingelagertem Rauschen

8

AURISCON BERÄT. PRÜFT. SICHERT.

Evasion-Angriff (Verkehrsschilder)

- R=(B,Ss,A,Sd,Z)
- *B: Manipulation der Eingabemenge, Ss: Algorithmus mit fehlerhaften Zuordnungen, A: Autonomes Fahrzeug, Sd: Falsche Reaktion des Fahrzeugs, Z: Integrität*

Achtung, das ist kein typisches Informationssicherheitsszenario mehr!

Informationssicherheit: Information Security

Physische Sicherheit: Safety

→Kombiniertes Szenario Security+Safety

Gegenmaßnahme: adversarielles Retraining, fokussiertes Tranings auf „berauschte“ Daten

9

AURISCON BERÄT. PRÜFT. SICHERT.

Noch ein KI-spezifisches Risiko

Bias und Diskriminierung: KI-Systeme können ungewollte Vorurteile und Diskriminierung verstärken, wenn sie auf verzerrten Daten trainiert werden.

- R=(B,Ss,A,Sd,Z)
- *B: Manipulation der Eingabemenge, Ss: Algorithmus mit mangelhafter Erkennung, A: ChatGPT, Sd: Beeinträchtigung der Persönlichkeitsrechte von Menschen, Z: Integrität*

Gegenmaßnahme: hochwertige und vielfältige Datensätze verwenden, die die gesamte Zielgruppe repräsentieren usw.

→ OWASP Top 10 for LLM Applications 2025

10

AURISCON BERÄT. PRÜFT. SICHERT.

Managementsystem für Künstliche Intelligenz

ISO/IEC 42001

ISO/IEC-Norm 42001 (2023)

Information technology — Artificial intelligence — Management system

6 **Planning** **8**
6.1 Actions to address risks and opportunities 8
6.1.1 General 8
6.1.2 AI risk assessment 9
6.1.3 AI risk treatment 9
6.1.4 AI system impact assessment 10
6.2 AI objectives and planning to achieve them 10
6.3 Planning of changes 11

Abschnitt 6.1 behandelt die Maßnahmen zur Bewältigung von KI-Risiken nach entsprechender Risikobeurteilung.

Risikobehandlung: Entwicklung und Implementierung geeigneter Strategien zur Minderung oder Beseitigung der identifizierten Risiken. Dies kann die Implementierung von Schutzmaßnahmen, regelmäßige Audits oder Schulungen für Benutzer der KI-Systeme umfassen.

11

AURISCON BERÄT. PRÜFT. SICHERT.

Managementsystem für Künstliche Intelligenz

ISO/IEC-Norm 42001 (2023) hat eine Ergänzungsnorm:
ISO/IEC 23894:2023
Informationstechnik – künstliche Intelligenz – Leitlinien für Risikomanagement
Auszug aus der Gliederung

4. Grundsätze des KI-Risikomanagements: Darstellung der fundamentalen Prinzipien für das Management von Risiken im KI-Bereich.
5. Rahmenwerk: Beschreibung der Struktur und Organisation des Risikomanagements
6. Prozesse: Detaillierte Beschreibung der Risikomanagementprozesse

Zusätzlich enthält die Norm informative Anhänge:

- Anhang A: Typische KI-bezogene Ziele.
- Anhang B: Risikoursachen im Zusammenhang mit KI.
- Anhang C: Risikomanagement und der Lebenszyklus von KI-Systemen.

→ Verantwortlichkeit, Fairness …

12