IFAF Institut für angewandte Forschung Berlin

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

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

Herausforderungen API-basierter KI-Dienste

(Blackbox, Vorhersagbarkeit, XAI, Robustness, Fairness, Compliance, …)
Impuls zum Workshop (KI-Sicherheit …) im Kontext des Projekts TAHAI,
HTW Berlin, 13. März 2025

Prof. Dr.-Ing. habil. Andreas Schmietendorf

HWR Berlin/FB2 sowie Otto-von-Guericke Universität Magdeburg/FIN

Fraunhofer IESE

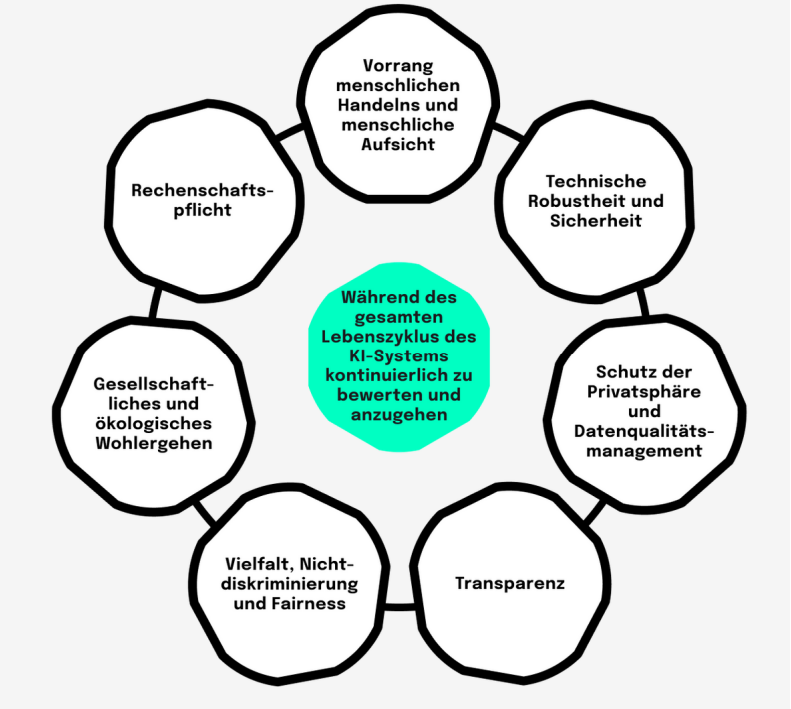
GESELLSCHAFT FÜR INFORMATIK

CMG Central Europe Computer Measurement Group

Projekt TAHAI – Vertrauenswürdigkeit

TrustAdHocAI

- Prototypische Implementierungen
  - Forstwirtschaft – Totholzerkennung
  - Professionsforschung – Reife von Mediationen
  - Bahninfrastruktur – Vandalismusbekämpfung
- Agile Umsetzung unter Verwendung von
  - Via KI-(Web) APIs bezogene KI-Modelle (LLMs, Vision, …)
  - Low-Code orientiertes Vorgehen
  - KI-gestütztes Software Engineering



Quelle der rechten Abb.: Bastian, N.: Vertrauenswürdige KI im Finanzsektor: Ethik als Wettbewerbsvorteil, https://neosfer.de/vertrauenswuerdige-ki, 26.02.2024

KI-Sicherheit

Arten von KI-Risiken:

- Verzerrung und Fairness
- Unzureichender Datenschutz
- Gefahr des Kontrollverlusts
- Existenzielle Risiken
- Böswilliger Missbrauch
- Cybersicherheit

KI-Sicherheitsmaßnahmen:

- Erkennung und Beseitigung von algorithmischer Verzerrung
- Testen und Validierung der Robustheit
- Explainable AI (XAI)
- Ethische KI-Frameworks
- Menschliche Überwachung
- Sicherheitsprotokolle
- Branchenweite Zusammenarbeit

In Anlehnung an: McGrath, A.; Jonker, A.: Was ist KI-Sicherheit?, 15.11.2024
https://www.ibm.com/de-de/think/topics/ai-safety

KI-Sicherheit – aktuelle Angriffsvektoren

- LLM01:2025 Prompt-Injektion
  (Unbeabsichtigte Änderungen am KI-Modell)
- LLM02:2025 Sensitive Information Disclosure
  (Offenlegung sensibler bzw. vertraulicher Daten)
- LLM03:2025 Supply Chain
  (Datenverzerrungen, Sicherheitsverletzungen, …)
- LLM04:2025 Data and Model Poison
  (Datenverfälschungen – ggf. vor dem Training)
- LLM05:2025 Improper Output Handling
  (Unsachgemäße Handhabung von Ausgaben)

Quelle: OWASP Top 10 for LLM Applications 2025, 17.11.2024, https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

APIs und KI-Sicherheit

„APIs (Application Programming Interfaces oder Programmierschnittstellen) sind nötig, um KI-Modelle zu trainieren und bereitzustellen. Doch bilden sie gleichzeitig auch mögliche Einfallstore für Cyberkriminelle, die Daten stehlen, sich unerlaubten Zugang verschaffen und KI-Modelle manipulieren. KI ist damit nur so sicher wie die Schnittstelle, über die sie bedient wird.“

Quelle des Zitats: Stephan Schulz: API-Schutz ist der erste Schritt zur KI-Sicherheit, 04.03.2025
https://www.bigdata-insider.de/ki-sicherheit-schutz-apis-a-25c2b8e6ff10be6835b85a39fa91c440

Beispiel OpenAI API – Objekterkennung in Bildern

- Zwingende Authentifizierung mit Hilfe eines API-Key (Bearer Token)
- Einsatz verschiedener LLMs bzw. ANNs (u.a. GPT-4o, DALL·E 2/3, Whisper, …)
- Build-In-Tools (Web-Search, File Search, Function calling, …)
- Computer-using agents (Algorithmische Steuerung und Simulation)
- …

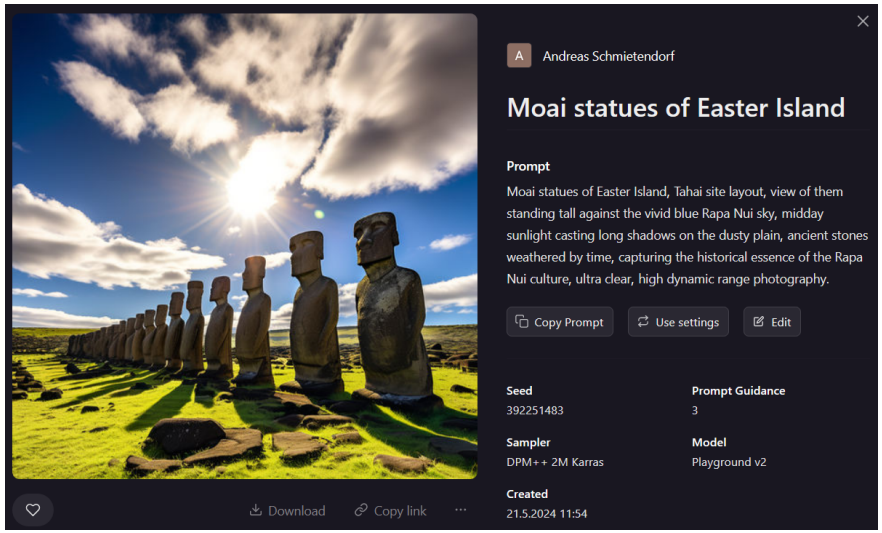
Analyze the content of an image — python

```python
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
    model="gpt-4o-mini",
    messages=[{
        "role": "user",
        "content": [
            {"type": "text", "text": "What's in this image?"},
            {
                "type": "image_url",
                "image_url": {
                    "url": "https://upload.wikimedia.org/wikipedia/commons/thumb/d/dd/Gfp-wisc
                },
            },
        ],
    }],
)

print(response.choices[0].message.content)
```

Quelle: OpenAI - Images and vision, https://platform.openai.com/docs/guides/images?api-mode=chat&lang=python

TAHAI-Erkenntnisse – High Level

- Abwägung KI-APIs (On-Premise) vs. KI-Web-APIs
- Bedarf anonymisierter Daten – hoher Aufwand!
- Einsatz von High- und Low-Code Ansätze
- Erforderliche Testansätze zur KI-Robustheit
- Qualität der Quelldaten als Erfolgskriterium
- Risiko-orientiertes Vorgehen – vgl. EU AI Act
- Agiles- bzw. Feedback gesteuertes Vorgehen
- ...



rechte Abbildung - KI-basiert erzeugt: https://playgroundai.com, 21. Mai 2024

Pragmatische Unterstützung

Bezugsbereich API-Security

- Safety – Funktionssicherheit

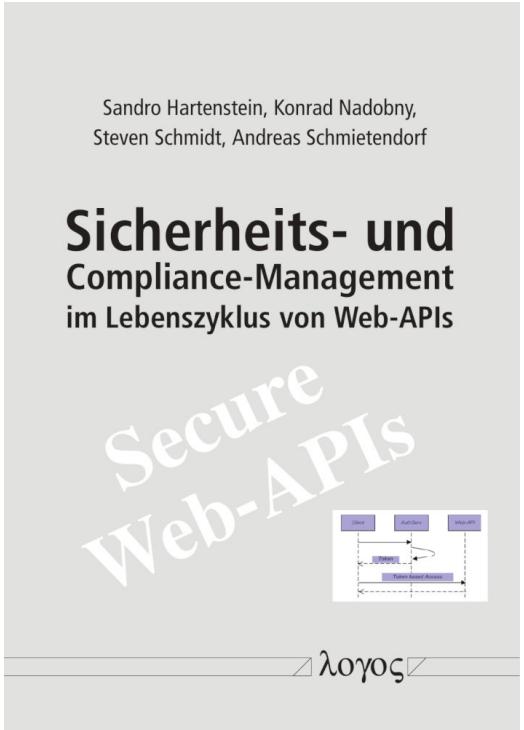
  (Funktional und Nicht-Funktional – u.a. Open API)

- Security – Informationssicherheit

  (Vertraulichkeit und Integrität – u.a. TLS, OAuth2)

- Compliance - Gesetze, Normen und Regeln

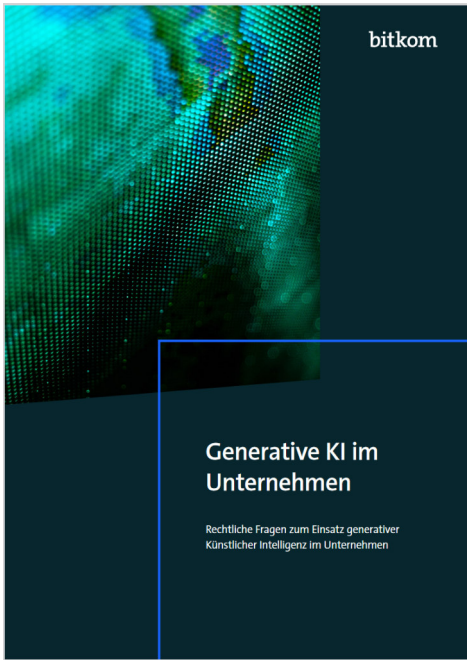
  (rechtliche Verpflichtungen – u.a. EU DSGVO)



Quelle: Hartenstein, S.; Nadobny, K.; Schmidt, S.; Schmietendorf, A.: Sicherheits- und Compliance-Management im Lebenszyklus von Web APIs, Ergebnisse eines Forschungsprojektes an der HWR Berlin/Uni Magdeburg, 140 Seiten, Monografie, Logos-Verlag, Berlin, März 2020

Bitkom Leitfäden

- Datenschutz (Rechte und Pflichten)
- IT-Sicherheit
  - Vergiftete Trainingsdaten (Data Poisoning)
  - Eingabeangriffe beim Promting
  - Modell-Extration (Reverse Engineering)
  - Denial-of-Service (DoS) - Anfragflut
- Schutzmaßnahmen – wie z.B.:
  - Robustes Training
  - Zugriffskontrollen und Verschlüsselung
  - Anomalieprüfung aller Eingabedaten
  - Rate Limiting

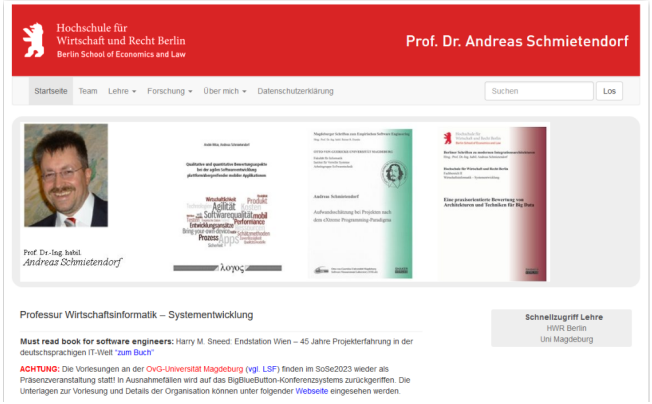


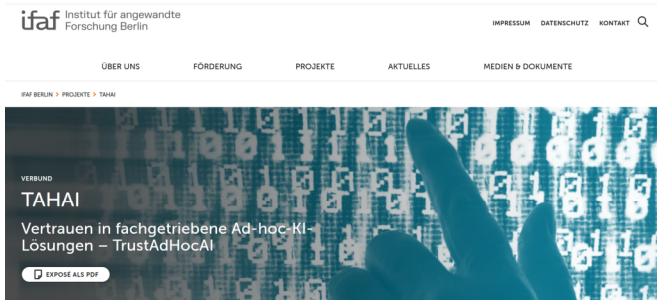
Quelle linke Abb.: https://www.bitkom.org/sites/main/files/2024-02/Bitkom-Leitfaden-Generative-KI-im-Unternehmen.pdf
Quelle rechte Abb.: https://www.bitkom.org/sites/main/files/2024-07/202407bitkom-leitfaden-ki-datenschutz.pdf

Weitere Informationen

Weitere Informationen

Schmietendorf, A.; Rodner, E.; Schnieders, R.: Herausforderungen Low-Code orientierter KI-Ansätze – Ergebnisse eines öffentlichen Expertenworkshops am Fraunhofer IESE unter Berücksichtigung der TAHAI-Projektergebnisse, in Berliner Schriften zu modernen Integrationsarchitekturen, Shaker-Verlag, Düren, Band 30, ISBN 978-3-8440-9729-0 (in Vorbereitung für I/2025)





Quelle der rechten Abb.: https://blog.hwr-berlin.de/schmietendorf, abgerufen 16.11.2023