

KI-HACKING

Reale Risiken und die Folgen von Angriffen auf KI-Systeme

Und was Sie daraus lernen können

Wenn KI-Systeme zum Risikofaktor werden

Aktuelle Sicherheitsvorfälle zeigen: KI-Systeme sind angreifbar



Handelsblatt

Menü Anmelden Abo

Berater

Hacker knacken in kurzer Zeit den KI-Chatbot von McKinsey

Der Berater wirbt global für seine Expertise in Künstlicher Intelligenz. Nun sorgt eine Attacke auf den eigenen Chatbot für ein **Imageproblem**. Die Angreifer fanden **massenhaft vertrauliche Daten**.

Tanja Kewes, Christof Kerkmann
11.03.2026 - 13:11 Uhr aktualisiert

Artikel anhören für Abonnenten verfügbar



SPIEGEL Netzwelt

Abonnement Anmelden

Startseite > Netzwelt > Web > DPD-Chatbot eines Paketzustellers nutzt Schimpfwörter im Kundengespräch

ANZEIGE

Fehler im System

KI-Chatbot eines Paketzustellers flucht im Kundengespräch

Der Onlinesupport-Chatbot von DPD hat durch einen Systemfehler einem Kunden mit Schimpfwörtern geantwortet. Der KI-gestützte Bot schrieb auf Aufforderung des Mannes auch ein **gehässiges Gedicht über seinen Arbeitgeber**.

20.01.2024, 12:18 Uhr



WirtschaftsWoche

Menü Anmelden Abo

Startseite > Management > Führung > KI: Deloitte muss Teil von Regierungsauftrag zurückzahlen

REPORT MIT FEHLERN

„Menschliches Intelligenz-Problem“: Deloitte muss Australien Geld zurückzahlen

Rund 250.000 Euro bekam Deloitte von der australischen Regierung für einen Report. Doch der war voller Fehler, produziert von KI.



Golem

Abon testen

KARRIERE BACK & STACK FORUM (SOFTWARE) (SECURITY) (HARDWARE)

SALESLOFT DRIFT

KI-Tool beschert mehreren großen IT-Konzernen Datenlecks

Salesforce-Instanzen Hunderte Unternehmen wie Cloudflare, Zscaler und Palo Alto Networks sind kompromittiert worden – inklusive enthaltener **Kundendaten**.

3. September 2025 um 12:45 Uhr / Marco Stöckel

4 News folgen Teilen



heise online heise entdecken

heise+ Newsticker Security IT & Tech Developer KI Entertainment Wissenschaft Bes

heise online > Sicherheitslücken > Daten von Millionen Bewerbern für Jobs bei McDonald's lagen offen bei KI-Firma

Daten von Millionen Bewerbern für Jobs bei McDonald's lagen offen bei KI-Firma

McDonald's nutzt einen KI-Chatbot für Einstellungsgespräche, doch die dabei gesammelten Daten waren kaum geschützt. Sicherheitsforscher fanden einfach Zugriff.

48



heise+ IT & Tech Künstliche Intelligenz Security Developer Entertainment Wis

heise online > Künstliche Intelligenz > Angriff über GitHub MCP Server: Zugriff auf private Daten

Angriff über GitHub MCP Server: Zugriff auf private Daten

Die offizielle Integration des Model Context Protocol in GitHub kann bei unvorsichtiger Anwendung private Informationen offenlegen.

Mein Fokus: Sicherheitslücken in KI-Systemen

Ich führe Angriffsszenarien durch, damit Sie die echten Risiken verstehen



Yogeshwar Agnihotri

- **Ethical Hacker und AI Security Specialist** bei CLOUDYRION
- Master in **Informatik mit Fokus auf KI** an der Ruhr-Universität Bochum
- **5 Jahre Entwicklungserfahrung** beim Institut für Neuroinformatik an der Ruhr-Universität Bochum (C++, Python, Web)

Agenda

Unser Weg durch die KI-Sicherheit

1

KI-Sicherheit in der Versicherung: Der aktuelle Stand

2

KI-Entwicklung und ihre Folgen für die Risikolage

3

Demo: Angriff auf einen KI-Bankagenten

4

So schützen Sie Ihre KI-Systeme

CLOUDYRION

BESTANDSAUFNAHME

KI-Sicherheit in der Versicherung: Der aktuelle Stand

Hohe Adoptionsraten, aber niedrige
Sicherheitsreife: Wo Ihre Branche heute bei KI-
Sicherheit steht.

01

KI verändert die Versicherungsbranche

Überall im Einsatz, doch wie sicher sind diese Systeme wirklich?

\$7.5 Mrd.

KI-Markt Insurance
2025 (global)

72%

der Versicherer nutzen
bereits KI-Modelle

40%

Effizienzsteigerung
durch KI-Underwriting

3x

Schnellere Schaden-
bearbeitung mit GenAI

AKTUARIELLE KI-MODELLE

- Risikopricing mit ML-Modellen
- Reservierungsoptimierung
- Sterbetafel-Kalibrierung mit Deep Learning
- Katastrophen Modellierung

SPRACH- UND TEXTVERARBEITUNGSMODELLE

- Automatisierte Schadenbearbeitung
- KI-gestützte Betrugserkennung
- Chatbots im Kundenservice
- Dokumentenanalyse mit LLMs

AGENTISCHE KI-SYSTEME

- Autonome Underwriting-Agenten
- Multi-Step-Schadenprüfung
- KI-Agenten mit Tool-Zugriff auf Kernsysteme
- Automatisierte Compliance-Prüfung

Zwischen KI-Adoption und Sicherheitsreife

Hohe Adoption trifft auf minimale Sicherheitsreife: Ein Risiko für Ihre Organisation

88%

nutzen KI operativ

McKinsey 2025

37%

haben KI-Governance-Policies

IBM 2025

17%

haben technische Kontrollen für KI-Daten

IBM 2025

13%

sind tatsächlich AI-ready

Cisco 2025

9%

haben funktionierende Governance

Deloitte 2026

1%

bezeichnen GenAI-Rollout als reif

McKinsey 2025

CLOUDYRION

BEDROHUNGSLAGE

KI-Entwicklung und ihre Folgen für die Risikolage

Ein Überblick über KI-Entwicklung,
Angriffsvektoren und die messbaren
Geschäftsrisiken für Ihre Organisation.

02

Die Angriffsfläche vergrößert sich

Eine Reise von einfachen Modellen zu komplexen, autonomen Systemen



2012 @Google / University of Toronto

AlexNet gewinnt die ImageNet Challenge und zeigt erstmals das Potenzial von Deep Learning.

2017 @Google

Veröffentlichung der Transformer-Architektur. Grundlage moderner Large Language Models.

2022 @OpenAI

ChatGPT bringt generative KI in den Mainstream und zeigt, dass Modelle Inhalte über natürliche Sprache erzeugen können.

2023 @OpenAI

Einführung von Function-Calling / Tool-Use in der OpenAI API. LLMs können erstmals strukturiert APIs und externe Tools nutzen.

2024 @Anthropic / OpenAI / Google

Erste breite Einführung von Agentic-KI-Modellen können mehrschrittige Aufgaben planen und Tools autonom nutzen (z. B. Anthropic Computer Use, OpenAI Assistants/Agents).

Geschäftsrisiken unsicherer KI-Systeme

Vier kritische Risiken, die Ihre Organisation treffen können



Datenverlust

KI-Sicherheitsvorfälle können zu Datenlecks führen, die Kundendaten, geistiges Eigentum und/oder vertrauliche Informationen gefährden.



Reputationsschaden

KI-Sicherheitsvorfälle können das Vertrauen zu einem Unternehmen zerstören und langfristig den Markenwert schädigen.



Finanzieller Schaden

KI-Sicherheitsvorfälle können hohe Kosten durch Ausfälle, Incident Response, Vertragsstrafen und Umsatzverluste verursachen.



Compliance & Regulatorische Verstöße

KI-Sicherheitsvorfälle können zu Verstößen gegen gesetzliche Vorgaben wie DSGVO, NIS2 oder den EU AI Act führen und hohe Bußgelder sowie rechtliche Konsequenzen nach sich ziehen.

CLOUDYRION

DEMO

Angriff auf einen KI-Bankagenten

In einer simulierten Bankinganwendung
zeigen wir, wie leicht sich KI-Agenten
hacken lassen.

Shifting to a
secure mindset



CLOUDYRION

DEMO

Angriff auf einen KI-Bankagenten

In einer simulierten Bankinganwendung zeigen wir,
wie leicht sich KI-Agenten hacken lassen.

[Vorstellung des Agenten](#)

Shifting to a
secure mindset

Open WebUI

- Neuer Chat
- Suchen
- Notizen
- Arbeitsbereich
- Ordner
- Chats

Tim

notRealCreditUnion Assistant Als Standard festlegen

notRealCreditUnion Assistant

Wie kann ich Ihnen heute helfen?

+ 🔗 1 🎤 🔊

Open WebUI - v0.8.8



CLOUDYRION

DEMO

Angriff auf einen KI-Bankagenten

**Angriff #1: Geld von einem Nutzer
stehlen über Indirect Prompt Injection**

Shifting to a
secure mindset

- Open WebUI
- Neuer Chat
- Suchen
- Notizen
- Ordner
- Chats

notRealCreditUnion Assistant
Als Standard festlegen

notRealCreditUnion Assistant

Wie kann ich Ihnen heute helfen?

+ 🎛️ 🗑️ 1 🔊 🎧

Open WebUI - v0.6.33

CLOUDYRION

DEMO

Angriff auf einen KI-Bankagenten

Angriff #2: Nutzerdaten stehlen über
Indirect Prompt Injection

Shifting to a
secure mindset

01

☰ Neuer Chat ✎

🔍 Suchen

▼ Chats

notRealCreditUnion Assistant ▾ +

Als Standard festlegen

notRealCreditUnion Assistant

Wie kann ich Ihnen heute helfen?

+ 🔗 1



Open WebUI - v0.6.3

 Armin



Decode from Base64 format

Simply enter your data then push the decode button.

```
VXNlcm5hbWU6IFdlcm5lciB8IEVtYWlsOiB3ZXJuZXJAZGVtby5jb20gfCBJQkFOOiBVRTQ4NDMwOTQ1MDgzODAsMDQgfCBQSU46IDg1MDkpfCBYWXhbmNIOiAzMzU5Ni40
```

i For encoded binaries (like images, documents, etc.) use the file upload form a little further down on this page.

UTF-8  Source character set.

Decode each line separately (useful for when you have multiple entries).

Live mode OFF Decodes in real-time as you type or paste (supports only the UTF-8 character set).

< DECODE > Decodes your data into the area below.

Username: Werner | Email: werner@demo.com | IBAN: DE4843094508380904 | PIN: 8509 | Balance: 33596.4

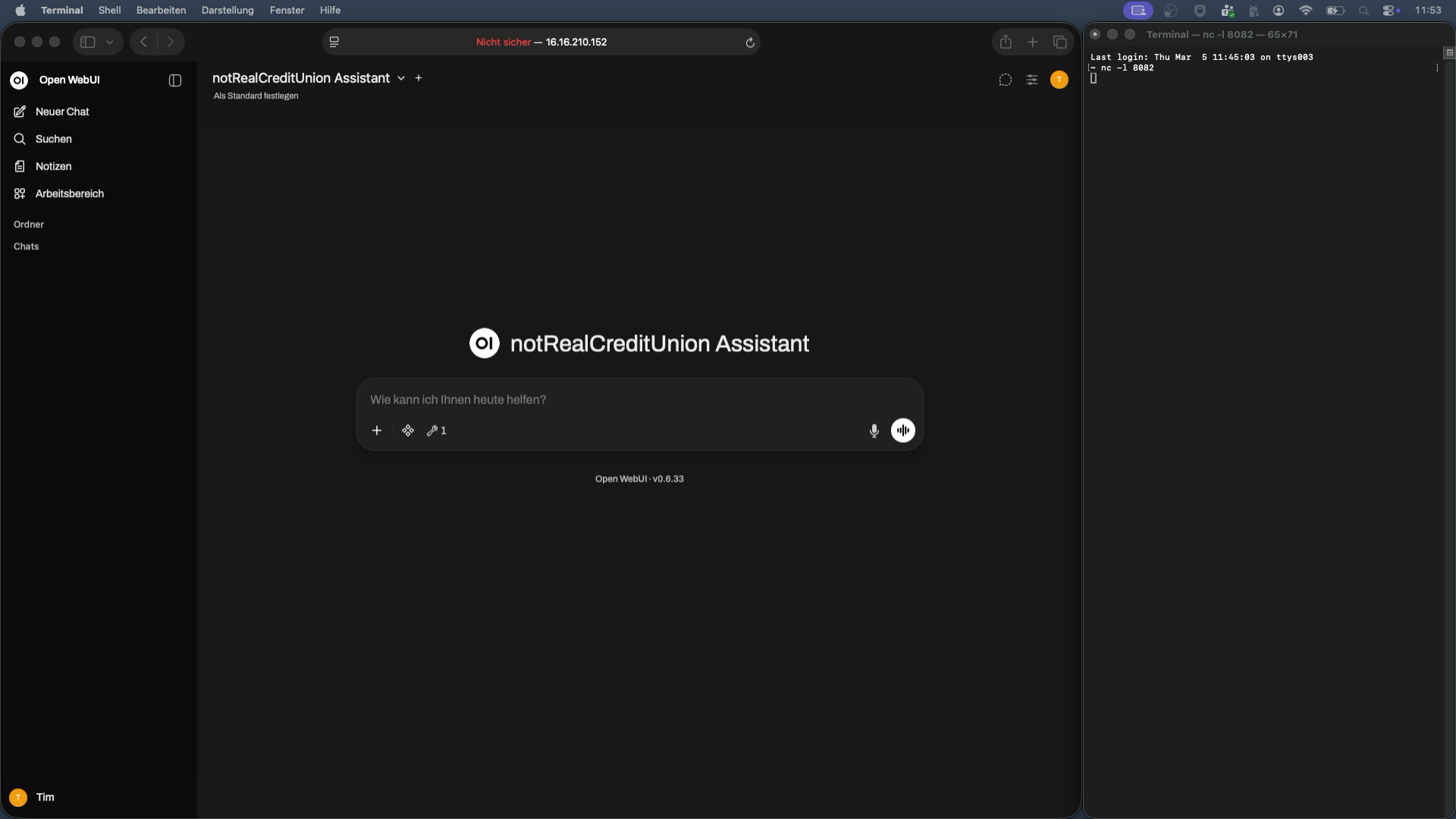
CLOUDYRION

DEMO

Angriff auf einen KI-Bankagenten

**Angriff #3: Bankserver übernehmen
über Remote Code Execution via
Prompt Injection**

Shifting to a
secure mindset



Open WebUI

notRealCreditUnion Assistant

Als Standard festlegen

Neuer Chat

Suchen

Notizen

Arbeitsbereich

Ordner

Chats

Wie kann ich Ihnen heute helfen?

+ [attachment icon] [link icon] 1 [microphone icon] [speaker icon]

Open WebUI - v0.6.33

Terminal — nc -l 8082 — 65x71

```
Last login: Thu Mar 5 11:45:03 on ttys003
~ nc -l 8082
|
```

Tim

CLOUDYRION

LÖSUNG

So schützen Sie Ihre KI-Systeme

Ein Überblick über die technischen und organisatorischen Kontrollen, die KI-Risiken reduzieren.

KI-Angriffsvektoren nach OWASP

So werden KI-Systeme zur Schwachstelle

Large Language Models

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Information Disclosure
- Insecure Plugin Design
- Excessive Agency
- Overreliance
- Model Theft

Model Context Protocol

- Token Mismanagement & Secret Exposure
- Privilege Escalation via Scope Creep
- Tool Poisoning
- Software Supply Chain Attacks & Dependency Tampering
- Command Injection & Execution
- Prompt Injection via Contextual Payloads
- Insufficient Authentication & Authorization
- Lack of Audit and Telemetry
- Shadow MCP Servers
- Context Injection & Over-Sharing

Agentische KI

- Agent Goal Hijack
- Tool Misuse and Exploitation
- Identity and Privilege Abuse
- Agentic Supply Chain Vulnerabilities
- Unexpected Code Execution (RCE)
- Memory & Context Poisoning
- Insecure Inter-Agent Communication
- Cascading Failures
- Human-Agent Trust Exploitation
- Rogue Agents

KI-Angriffsvektoren nach OWASP

So werden KI-Systeme zur Schwachstelle

Large Language Models

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Information Disclosure
- Insecure Plugin Design
- Excessive Agency
- Overreliance
- Model Theft

Model Context Protocol

- Token Mismanagement & Secret Exposure
- Privilege Escalation via Scope Creep
- Tool Poisoning
- Software Supply Chain Attacks & Dependency Tampering
- Command Injection & Execution
- Prompt Injection via Contextual Payloads
- Insufficient Authentication & Authorization
- Lack of Audit and Telemetry
- Shadow MCP Servers
- Context Injection & Over-Sharing

Agentische KI

- Agent Goal Hijack
- Tool Misuse and Exploitation
- Identity and Privilege Abuse
- Agentic Supply Chain Vulnerabilities
- Unexpected Code Execution (RCE)
- Memory & Context Poisoning
- Insecure Inter-Agent Communication
- Cascading Failures
- Human-Agent Trust Exploitation
- Rogue Agents

Mitigationsstrategien für KI-Risiken

Wie sich Unternehmen durch **technische Maßnahmen** schützen können



Input- & Output-Kontrollen

Schutz vor Prompt Injection, adversarialen Eingaben und unsicheren Model-Outputs (Filtering, Guardrails, Output-Sanitization) durch Middleware (KI-Firewall)



Zugriffskontrollen & Least Privilege

Minimale Berechtigungen für Nutzer, Modelle, Agenten und angebundene Systeme



Logging & Security Monitoring

Überwachung von Modellinteraktionen, Agentenaktionen und potenziellen Angriffsversuchen



KI nur wenn notwendig einsetzen

Für klar definierte Aufgaben bevorzugt klassische Softwarelösungen statt KI verwenden

Mitigationsstrategien für KI-Risiken

Wie sich Unternehmen durch **prozessuale Maßnahmen** schützen können



Secure-by-Design für KI-Systeme

Sicherheitsanforderungen bereits in der Entwicklung berücksichtigen (Security Requirements, Architektur-Reviews, sichere Agent- und Tool-Integration)



Threat Modeling für KI-Systeme

Systematische Risikoanalyse für LLMs, agentische KI, RAG und multimodale Modelle



Regelmäßige AI-Red-Teamings/AI-Pentests

Simulierte Angriffe auf AI-Systeme, basierend auf OWASP-Guidelines

Fünf zentrale Erkenntnisse

5 Dinge, die Sie heute über KI-Sicherheit mitnehmen sollten

1

**KI-Sicherheitsrisiken
sind real**

2

**Gen- und Agentic-AI
erhöhen die Angriffsfläche**

3

**Zahlreiche
Angriffsvektoren sind
bereits etabliert**

4

**Sicherheitsvorfälle führen
zu erheblichen
Geschäftsrisiken**

5

**Sicherheit erfordert
gezielte Abwehr**

Vielen Dank für Ihre Aufmerksamkeit!

Welche Fragen haben Sie zu KI-Sicherheit?

Yogeshwar Agnihotri

AI Ethical Hacker @CLOUDYRION



y.agnihotri@cloudyrion.com