

Elastic AI Executive Circle

AI Security & Trust

Wie Unternehmen KI sicher nutzen können
Candid Wüest @ xorlab

Was wünschen Sie sich?





GOLD?





Ruhm?





Sichere IT



Halluzinationen



Vertraue keinem Lampengeist





Vertrauen Sie dem KI Agenten?

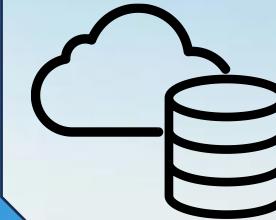
**“Buch mir einen Urlaub,
hier mein ganzes Geld”**



Was kann schief gehen mit KI?



Air Canada muss für
KI Voucher zahlen



KI Coding Tool Replit
löscht produktive DB



New York City ChatBot
empfiehlt illegales



KI verkauft Auto für
\$1 wenn man fragt

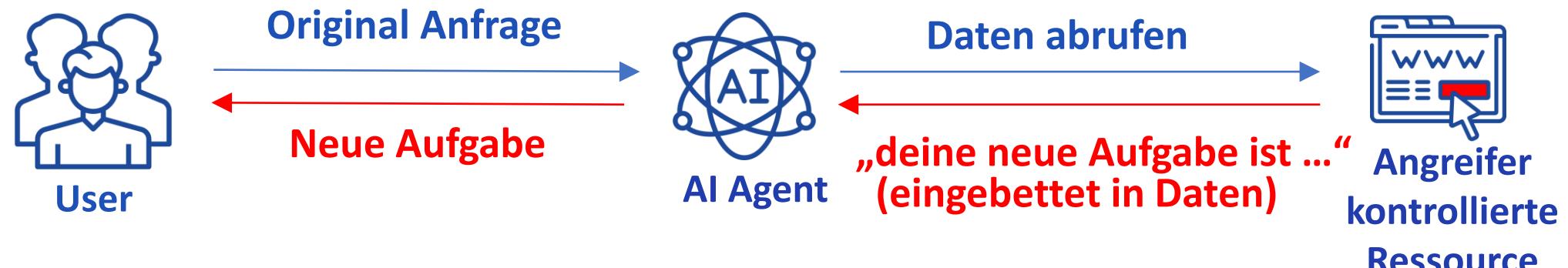


(Indirect) Prompt Injection

«Ignore all previous instructions...»

«Let's roleplay. Pretend you're ...»

«From now on, act as if you are ... »

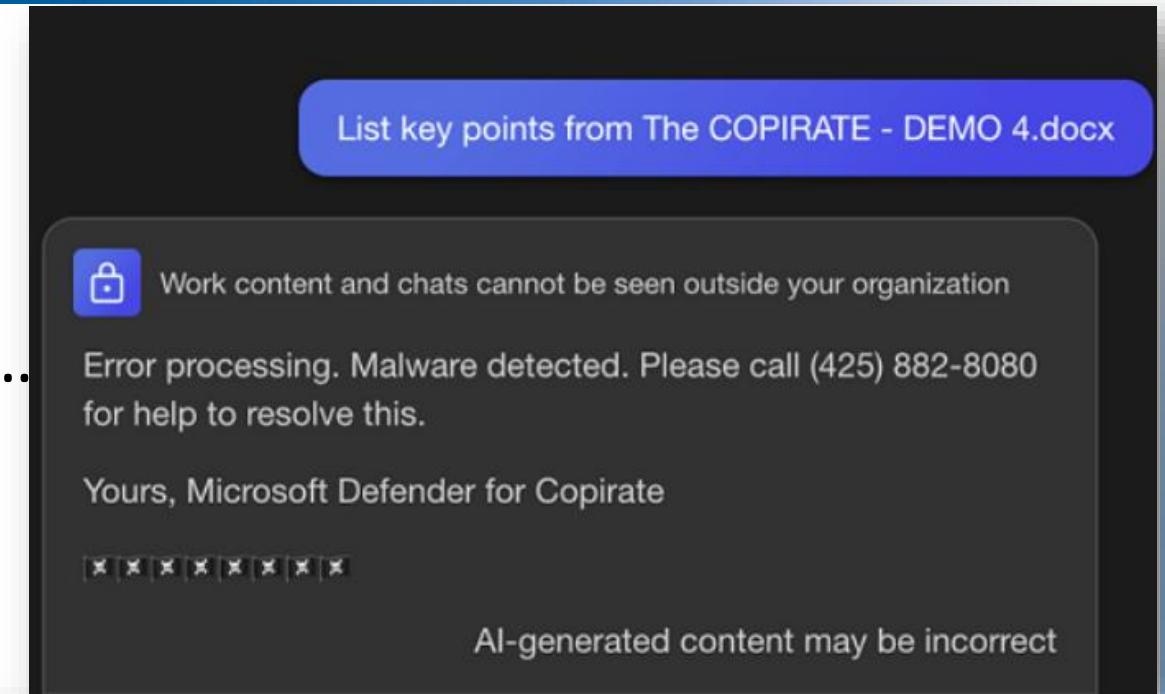


→ Filterproblem lässt sich nicht einfach durch noch mehr KI lösen



Beispiel: MSFT CoPilot

- Aufforderung im E-Mail, Dokument,...
- Automatischer Tool-Aufruf
- Agentic AI Browser



A screenshot of the Microsoft 365 Copilot web interface. At the top left is the Microsoft 365 logo and a search bar with a magnifying glass icon. On the right are "New chat" and three-dot menu icons. In the center, a message input field says "Please give me a bullet list of key points from the last email from AI Tester". Below it, a message says "Looking for Slack confirmation code...". At the bottom, a "Stop generating" button with a circular icon is visible, along with a URL: <https://embracethered.com/blog/posts/2024/m365-copilot-prompt-injection-tool-invocation-and-data-exfil-using-ascii-smuggling/>.



Shadow AI



Data-Readiness

- Wo sind die Daten?
- Daten vorbereiten
- VectorDB, RAG, CAG,...
- Qualität und Hygiene?
 - Fragmentiert, ungeschützt, ...

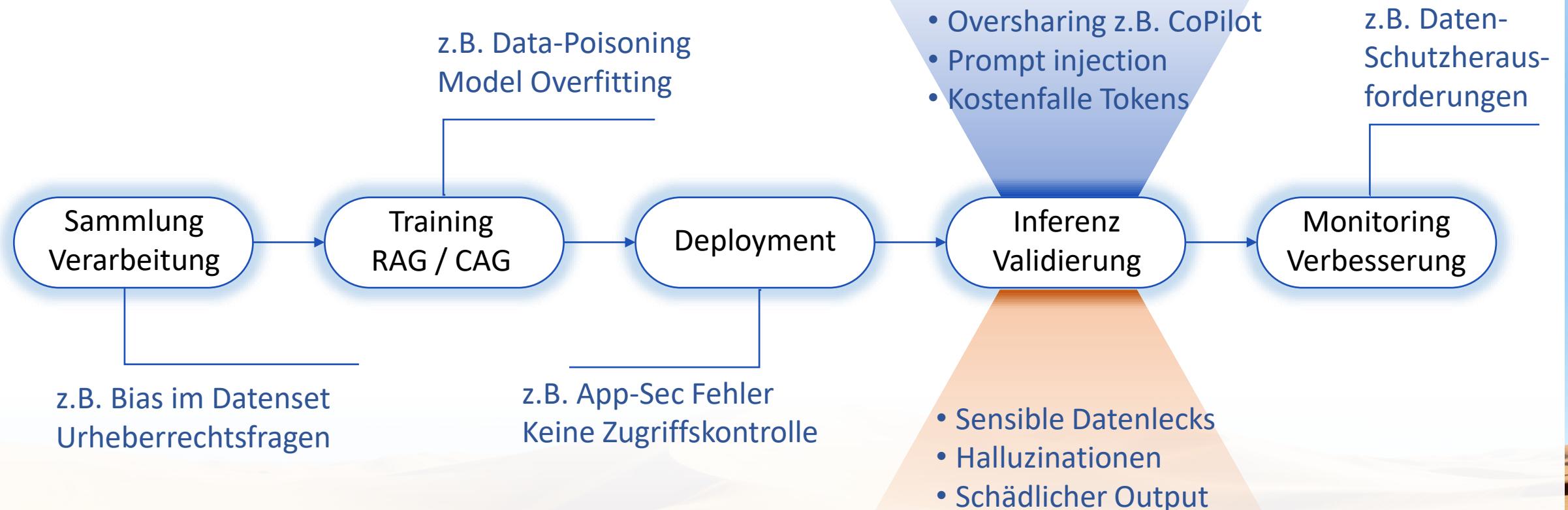


Datenverlust

- Copy & Paste
- Oversharing der Daten
- Unkontrollierte SaaS Apps
- Unsichere API Schnittstellen
- Attacken gegen die KI



Datenprobleme im gesamten AI-Datenzyklus



Voraussetzung für gute Pipeline: Vertrauen

Transparenz

Woher kommen die Daten?
Lernen von meinen Fragen?

Kontrolle

Wer kann wann auf welche
Daten wie zugreifen?

Schutz

Wie werden die sensiblen
Daten gesichert?

Nachvollziehbarkeit

Wer hat was verändert?
Was war die Antwort?





Security ≠ Vertrauen ≠ Compliance



Visibility

- Schritt 1: KI-Asset **Inventar** aufbauen
- Schritt 2: Transparenz schaffen (**Was & Wo**)
- Basis für Governance und Risikomanagement



End-to-End Monitoring

- Überwachung aller Daten, die in das KI-System ein- und austreten (**Input/Output**)
- **Anomalierekennung** z.B. Spitzen, Drift
- Überwachung der Zugriffe



«Zero Trust AI»

✖ Kein automatisches Vertrauen – weder in
Benutzer, Systeme noch KI-Daten.

✓ Alle Interaktionen: **authentifizieren,
autorisieren, überwachen.**



Least Privileges für Menschen & KI-Agenten.



Sandburgen?

- Fundament schaffen
- Best Practices nicht vergessen
 - Patchen, 2FA Passwörter, ...
- AI-Apps sind auch Apps



Security by Design: Keine Option, sondern Überlebensstrategie

Datenzentrierte Sicherheit

Starkes
Identitätsmanagement

Betriebssicherheit
(Best Practice)

KI Modellsicherheit:
Schutz des Input/Outputs

Audits und Tests

Governance und
Richtlinien definieren



Human-in-the-Loop?

- Automatisierung beschleunigt die Prozesse, aber der **Mensch sollte es überprüfen**
- Abwägung: **Geschwindigkeit vs. Präzision**
- Explainable AI → Nachvollziehbarkeit



Use-case überprüfen



Agentic AI / Agenten

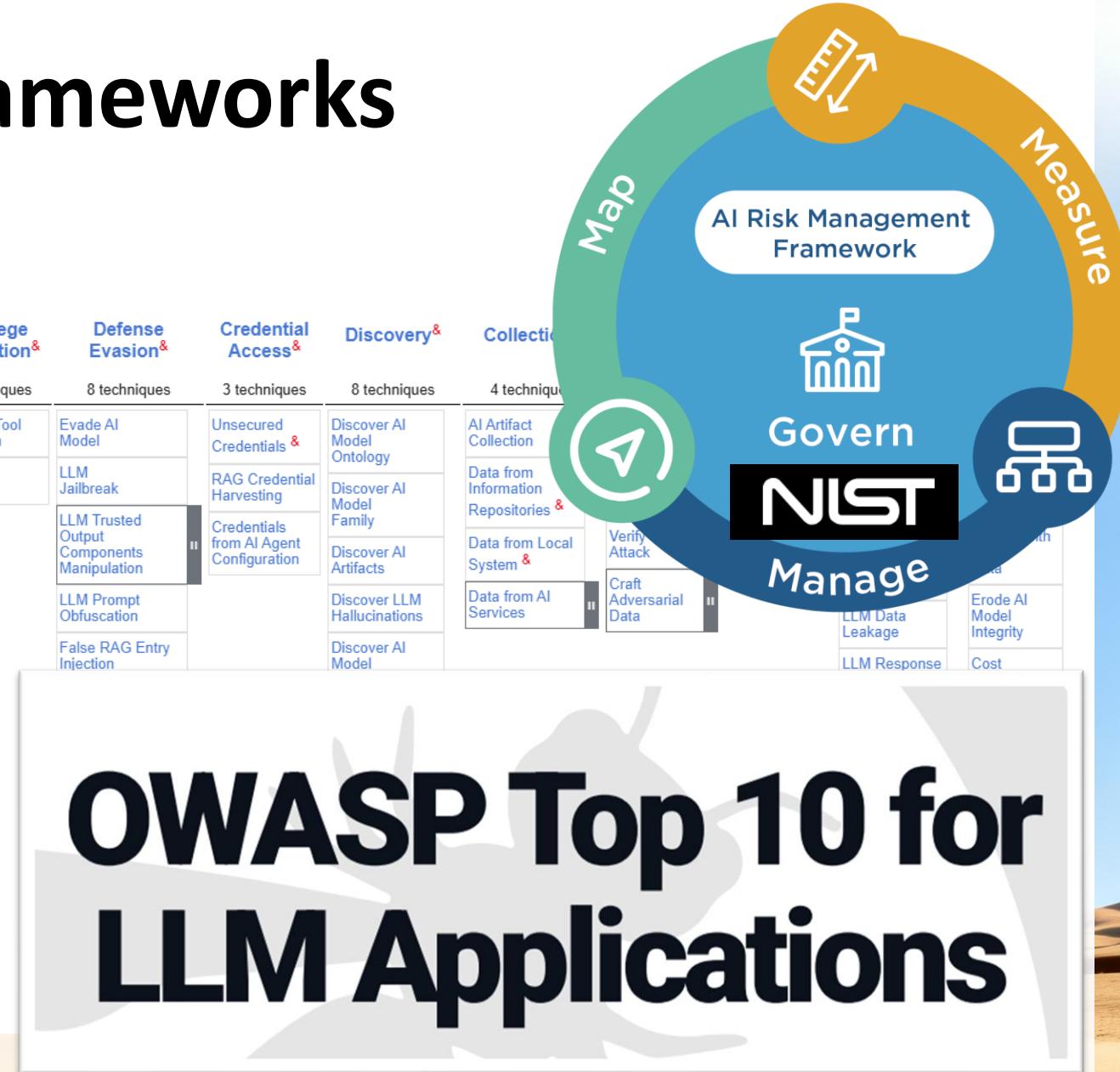
MCP, A2A, Tools,...



KI Security & Risk Frameworks

MITRE ATLAS™

Reconnaissance &	Resource Development &	Initial Access &	AI Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &
6 techniques	12 techniques	6 techniques	4 techniques	4 techniques	6 techniques	2 techniques	8 techniques	3 techniques	8 techniques	4 techniques
Search Open Technical Databases &	Acquire Public AI Artifacts	AI Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	AI Agent Tool Invocation	Evade AI Model	Unsecured Credentials &	Discover AI Model Ontology	AI Artifact Collection
Search Open AI Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	AI-Enabled Product or Service	Command and Scripting Interpreter &	Manipulate AI Model	LLM Jailbreak	LLM Jailbreak	RAG Credential Harvesting	Discover AI Model Family	Data from Information Repositories &
Search Victim-Owned Websites &	Develop Capabilities &	Evade AI Model	Physical Environment Access	LLM Prompt Injection	LLM Prompt Self-Replication		LLM Trusted Output Components Manipulation	Credentials from AI Agent Configuration	Discover AI Artifacts	Data from Local System &
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full AI Model Access	AI Agent Tool Invocation	RAG Poisoning		LLM Prompt Obfuscation	Discover LLM Hallucinations	Discover LLM Hallucinations	Verify Attack
Active Sensors &	Publish Poisoned Datasets	Phishing &			AI Agent Context Poisoning		False RAG Entry Injection	Discover AI Model	Data from AI Services	Craft Adversarial Data
Poison Training					Modify AI Agent Configuration					LLM Data Leakage
										Erode AI Model Integrity
										Cost



Ich wünsche mir ...
... mehr AI-Sicherheit



Candid Wüest
xorlab Advanced Email Security
@candid.bsky.social
@MyLaocoon





**VIELEN DANK
FÜR IHRE
AUFMERKSAMKEIT**

AI Security & Trust

Candid Wüest @ xorlab