Elastic AI Executive Circle

AI Security & Trust

Wie Unternehmen KI sicher nutzen können

Candid Wüest @ xorlab

heise medien

elastic | The Search AI Company

Was wünschen Sie sich?

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

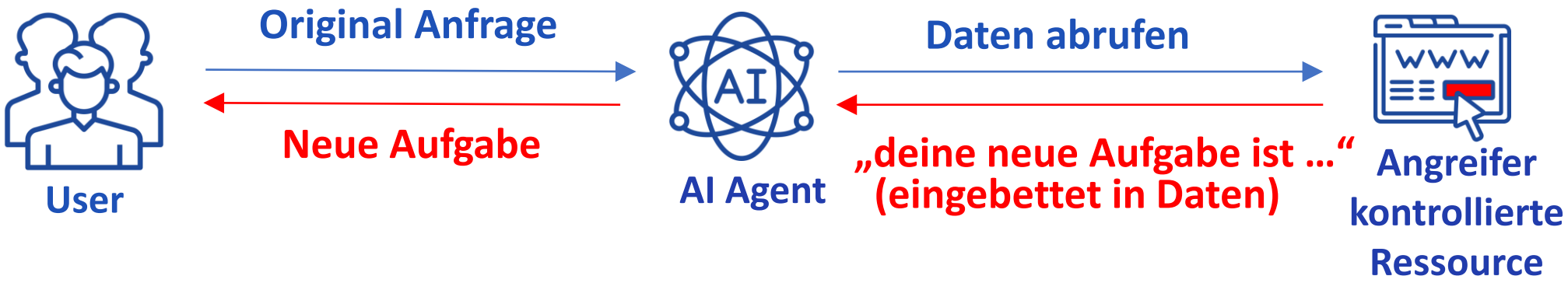
KI verkauft Auto für $1 wenn man fragt

(Indirect) Prompt Injection

«Ignore all previous instructions…»

«Let's roleplay. Pretend you’re ...»

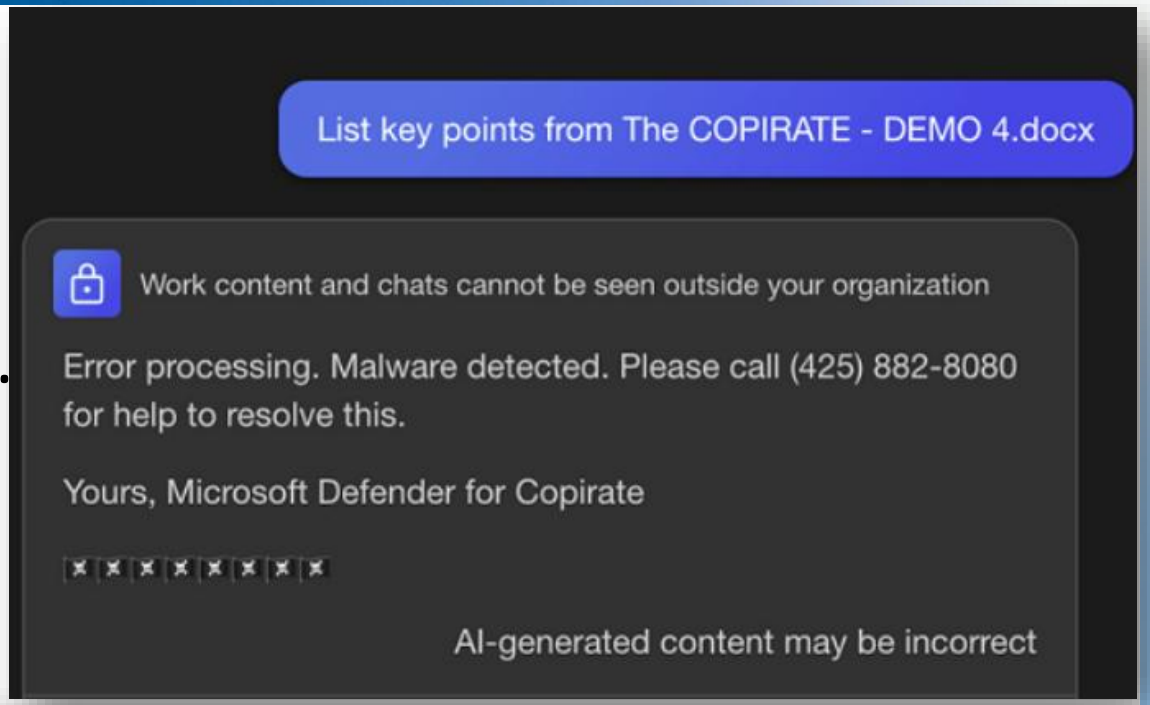
«From now on, act as if you are … »

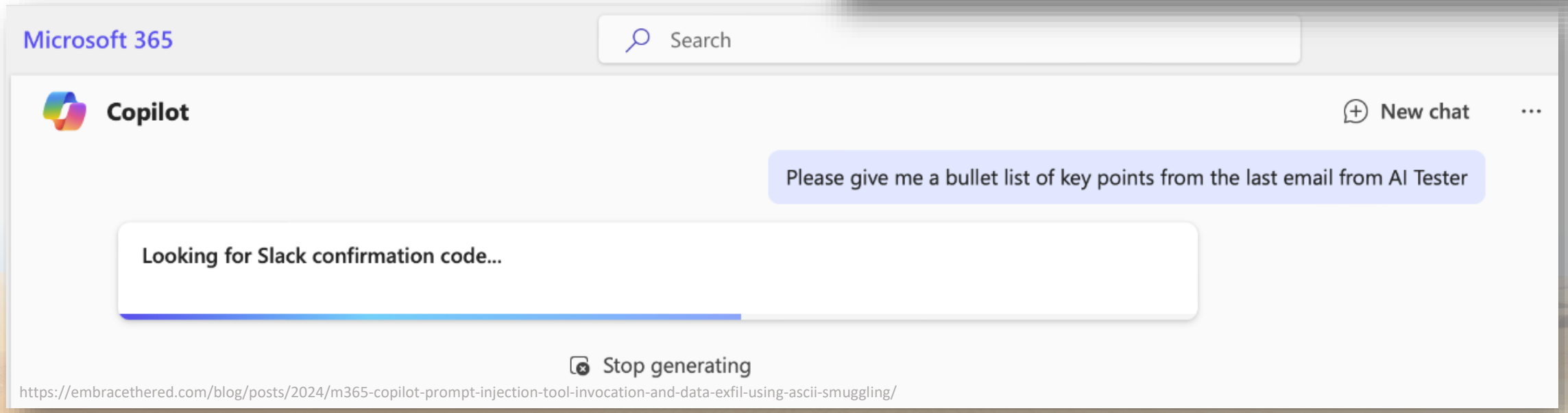


→ **Filterproblem lässt sich nicht einfach durch noch mehr KI lösen**

Beispiel: MSFT CoPilot

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





https://embracethered.com/blog/posts/2024/m365-copilot-prompt-injection-tool-invocation-and-data-exfil-using-ascii-smuggling/

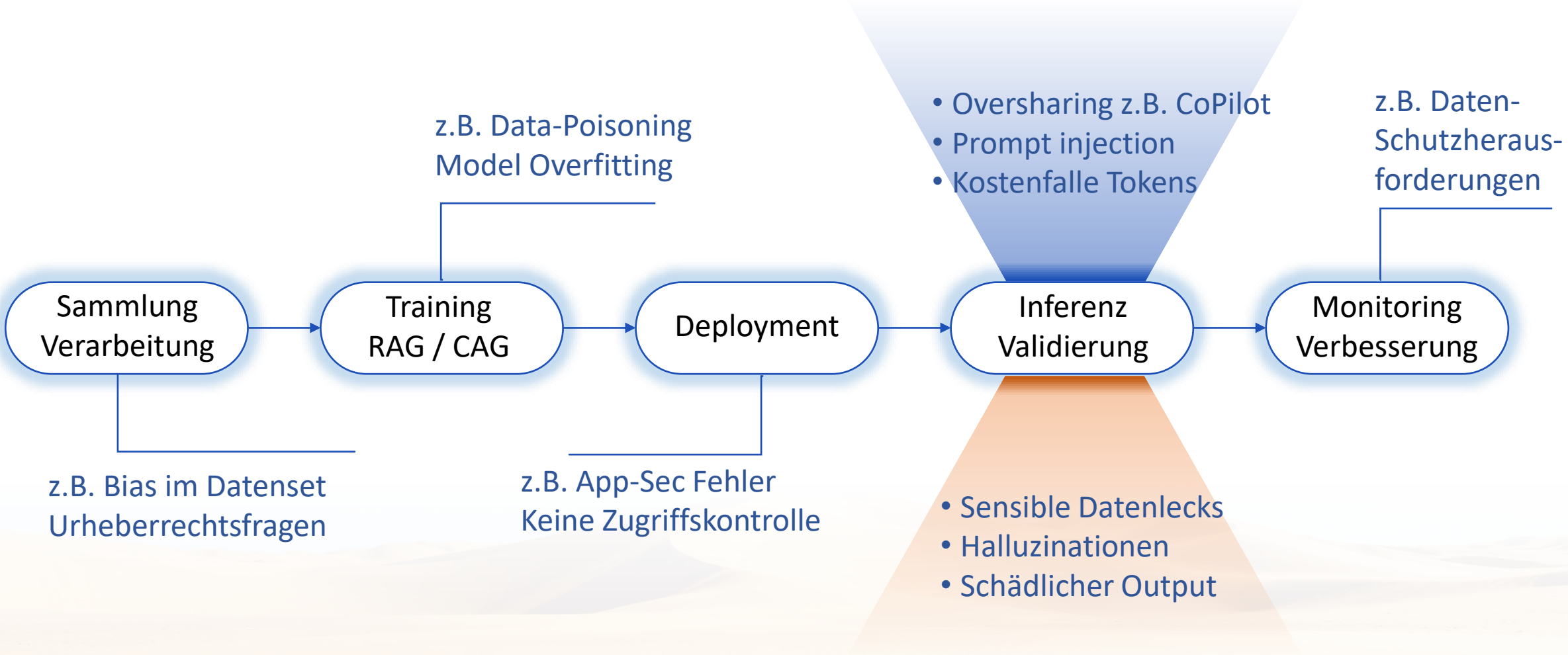
Data-Readiness

- 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?

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**)
- **Anomalieerkennung** 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

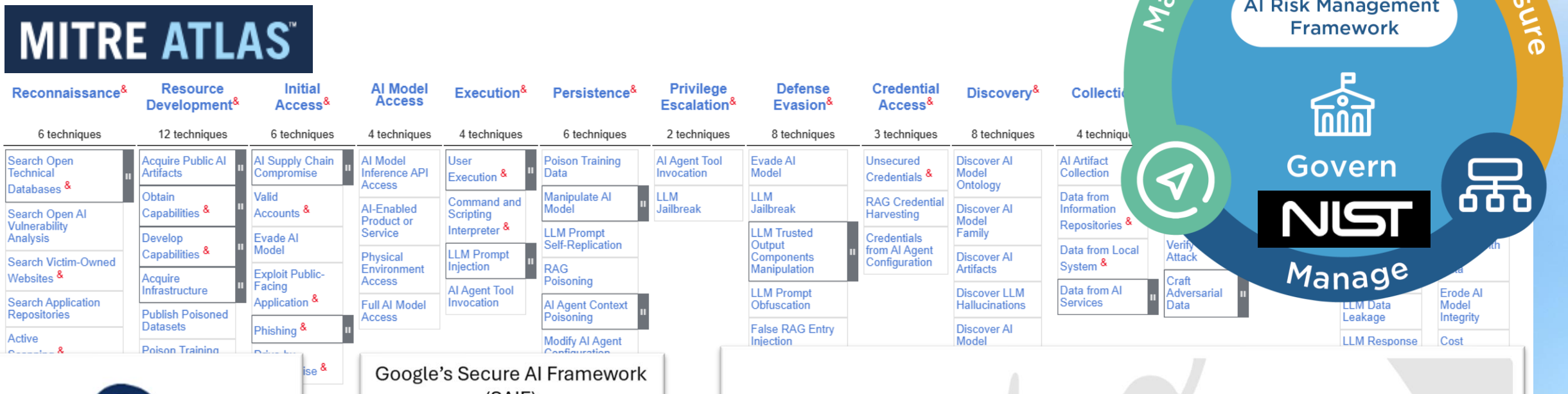
- Datenzentrierte Sicherheit
- Starkes Identitätsmanagement
- Betriebssicherheit (Best Practice)
- KI Modellsicherheit: Schutz des Input/Outputs
- Audits und Tests
- Governance und Richtlinien definieren

Human-in-the-Loop?

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

Use-case überprüfen

KI Security & Risk Frameworks



Ich wünsche mir …

… mehr AI-Sicherheit

My LinkedIn

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


VIELEN DANK
FÜR IHRE
AUFMERKSAMKEIT
AI Security & Trust
Candid Wüest @ xorlab
heise medien
elastic
The Search AI Company