Shadow AI
zu
Secure AI Operation

Wie MSPs sichere KI-Services anbieten können

CYBER SECURITY
neosec

Who is Sven?

Who is Sven?

Developer Advocate · Secure Coding · Java · AI/RAG Security

- Fokus: sichere Softwareentwicklung, Java und kontrollierbare KI-Systeme
- Praxisnähe: RAG, Self-Hosting, Datenschutz und Security Operations
- Ziel heute: KI nicht als Experiment, sondern als Managed Service denken

CYBER SECURITY neosec

1 · Das MSP-Problem

Shadow AI entsteht schneller als kontrollierte KI-Services.

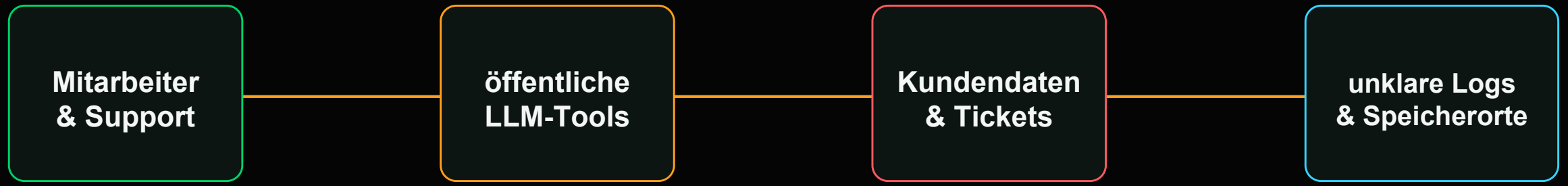
Warum das MSPs betrifft

KI wird Teil von Support, Betrieb, Dokumentation und Kundenkommunikation.

- MSPs verwalten sensible Kundendaten, Zugänge und Betriebsinformationen
- KI-Tools landen schnell in Ticketing, Wissensdatenbanken und Automatisierungen
- Ein Fehler skaliert nicht lokal — er kann mehrere Kundenumgebungen betreffen
- Daraus entsteht ein neues Managed-Security- und Compliance-Thema

Shadow AI im Kundenumfeld

Nicht freigegebene KI-Nutzung ist selten böse Absicht — meist ist sie schneller als Governance.



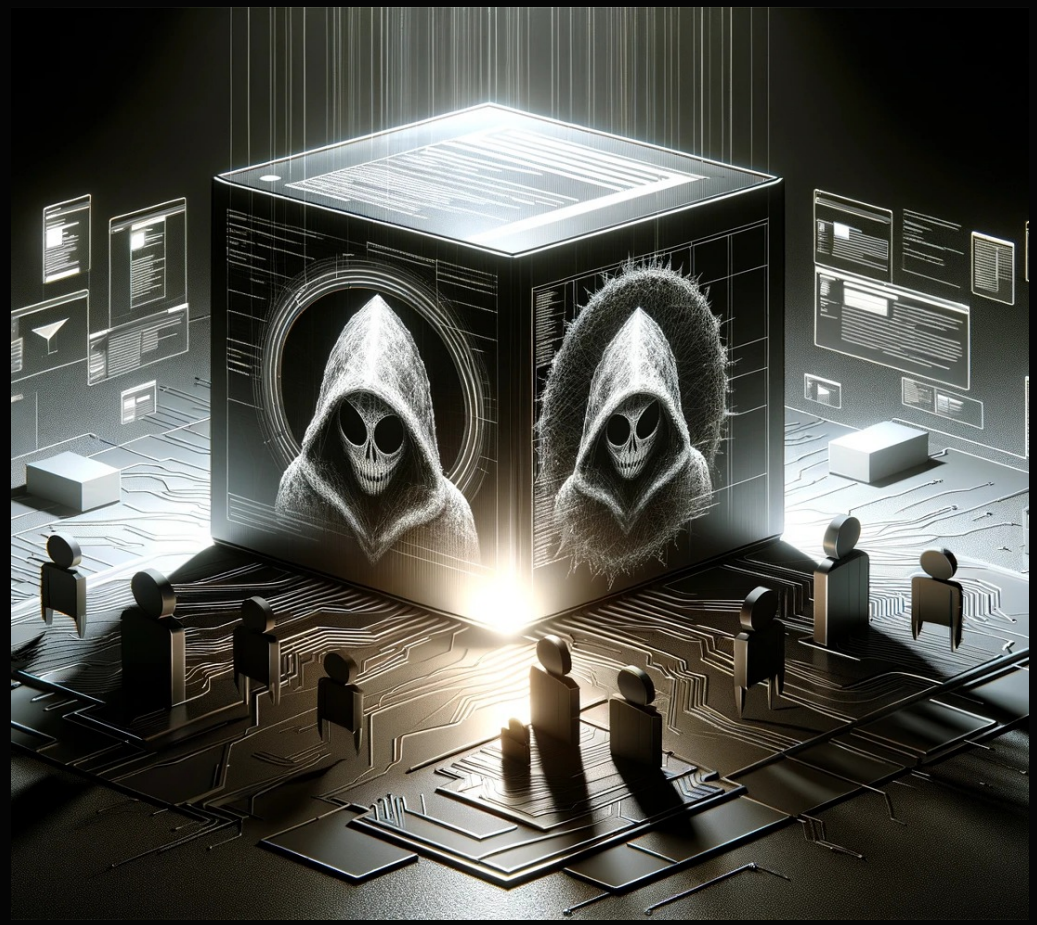
Typische Auslöser

- „Ich fasse nur schnell ein Kundenticket zusammen.“
- „Ich lasse mir eine Mail an den Kunden formulieren.“
- „Ich lade eine Konfiguration zur Fehlersuche hoch.“

Vom Tool-Risiko zum Betriebsrisiko

Sobald KI in MSP-Prozesse integriert wird, ist sie Teil der Angriffsfläche.

- Kundendaten fließen durch Prompts, Retrieval, Logs und Monitoring
- Automatisierung verstärkt Fehlverhalten und falsche Empfehlungen
- Connectoren, Plugins und Agents erweitern die Supply Chain
- Die zentrale Frage: Wer kontrolliert Daten, Modelle, Prompts und Outputs?



AI Supply Chain im MSP-Betrieb

Die KI-Lieferkette besteht nicht nur aus dem Modell.

- LLM-Provider, Embedding-Modell, VectorDB und Reranker
- Datenquellen: Tickets, Wikis, PDFs, E-Mails, Kundendokumente
- Integrationen: Browser-Plugins, SaaS-Connectoren, Agents, APIs
- Betrieb: Logs, Telemetrie, Backups, Admin-Zugänge und Policies

2 · Warum KI anders ist

Daten, Kontext und Modellverhalten sind enger gekoppelt als in klassischen Anwendungen.

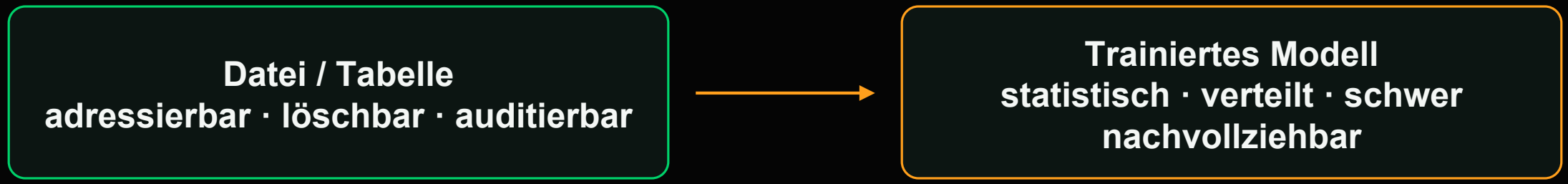
Datenhoheit: Kontrolle schlägt Speicherort

Der Standort allein beantwortet die Sicherheitsfrage nicht.

- Entscheidend ist, wer technische und rechtliche Kontrolle über Daten hat
- Subprozessoren, Parent Companies und Support-Zugriffe erzeugen indirekte Risiken
- MSPs müssen Mandantenfähigkeit, Schlüssel, Logs und Zugriffspfade prüfen
- „EU-Rechenzentrum“ ist kein Ersatz für Governance und Nachweisbarkeit

Ein LLM ist keine Datenbank

Gelöschte Dateien verschwinden — gelernte Muster nicht automatisch.



- Klassische Datenhaltung kennt Datensätze, IDs, Lösch- und Änderungsoperationen
- LLMs repräsentieren Wissen verteilt über Parameter und Wahrscheinlichkeiten
- Für MSPs heißt das: sensible Daten gehören nicht in unkontrollierte KI-Prozesse

RAG ist kein Datenschutz-Wundermittel

RAG reduziert Trainingsrisiken, erzeugt aber neue Betriebsrisiken.

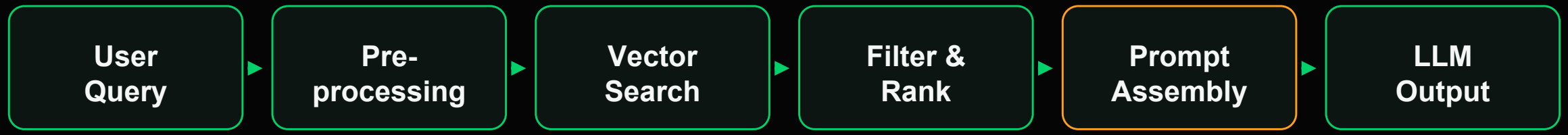
- Dokumente werden nicht zwingend trainiert, aber indexiert und wiederverwendet
- Embeddings, Metadaten und Snippets enthalten semantische Informationen
- Retrieval entscheidet, welche Kundendaten in den Prompt gelangen
- Sicherheit entsteht durch Pipeline-Kontrolle, nicht durch das Wort „RAG“

3 · RAG verstehen

Für MSPs zählt nicht jeder Algorithmus —
sondern jeder Kontrollpunkt.

RAG in einem Bild

Die Antwort entsteht aus Anfrage, Retrieval-Kontext und Modellgenerierung.

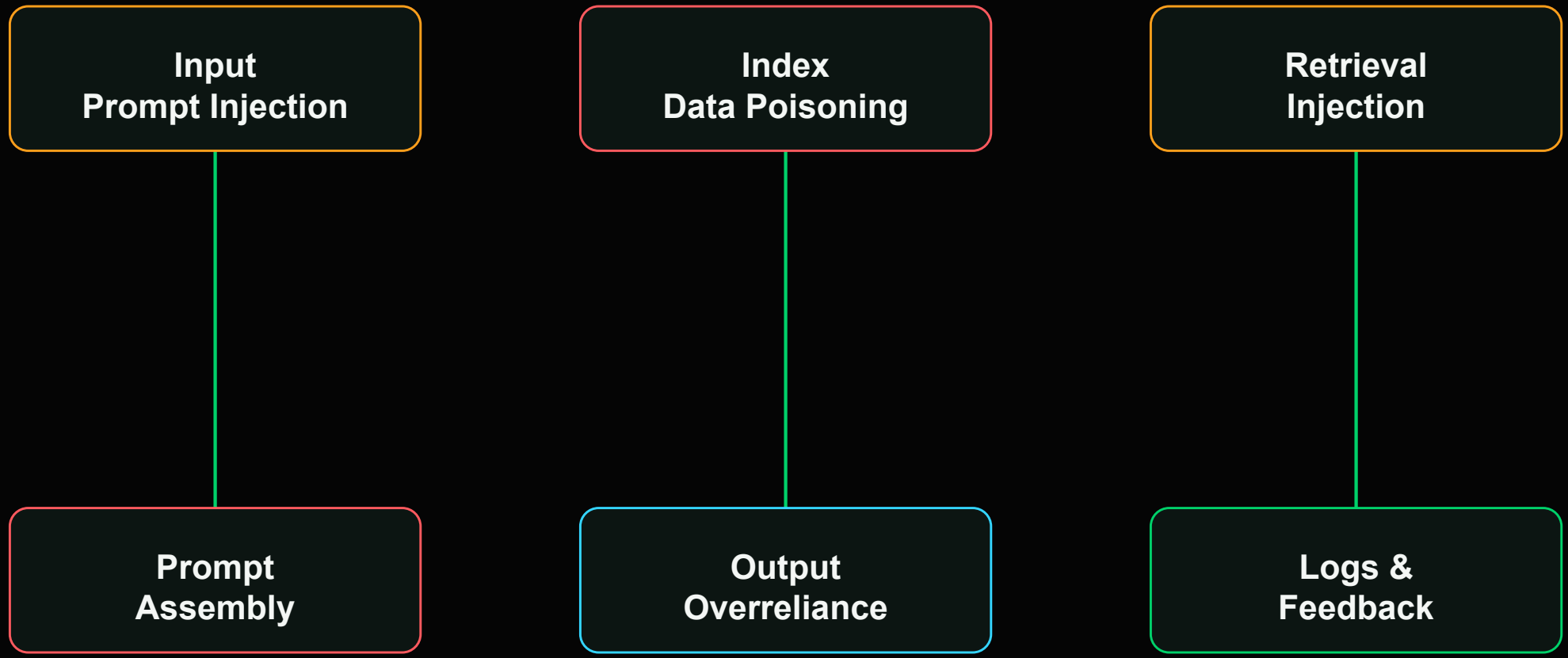


Kontrollpunkte für MSPs

- Eingaben klassifizieren und Mandant sauber bestimmen
- Datenquellen und Index-Inhalte validieren
- Zugriff, Trust-Level und Quarantäne vor Prompt Assembly erzwingen
- Output prüfen, zitieren, protokollieren und nachvollziehbar machen

Wo entstehen die Risiken?

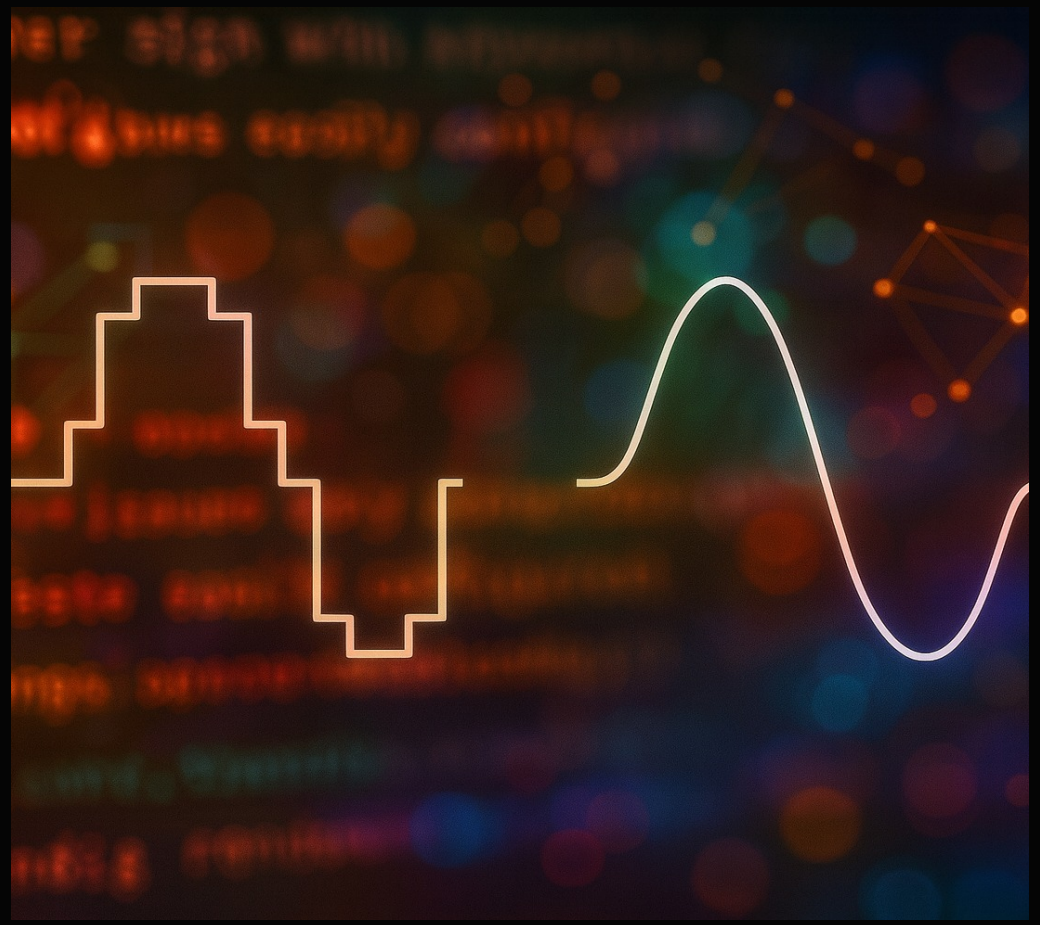
Angriffe können an mehreren Übergängen der RAG-Pipeline ansetzen.



Die VectorDB wird zum Schutzobjekt

Sie ist nicht nur ein technischer Index, sondern ein semantischer Zugriffspunkt.

- Embeddings können sensible Nähebeziehungen und Themencluster sichtbar machen
- Manipulierte Dokumente beeinflussen spätere Antworten indirekt
- Index-Versionen, Modellwechsel und Re-Embedding brauchen Governance
- MSP-Frage: Wer darf welche Inhalte indexieren, suchen und abrufen?



Access Control und Trust Metadata

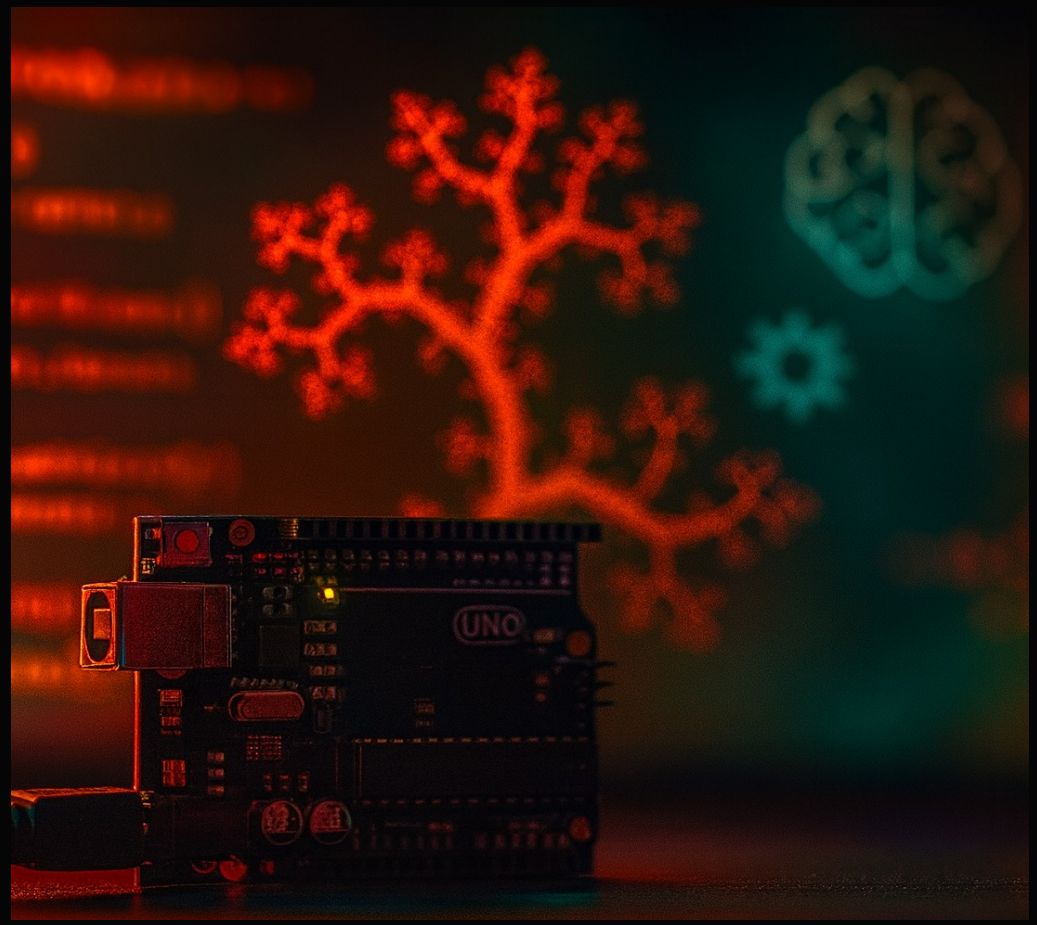
Retrieval darf Berechtigungen nicht umgehen.

- Berechtigungen müssen vor Context Construction geprüft werden
- Metadaten wie Kunde, Vertraulichkeit, Quelle und Aktualität sind sicherheitsrelevant
- Reranking darf nicht nur Relevanz, sondern auch Vertrauenswürdigkeit berücksichtigen
- Ergebnisqualität ohne Zugriffskontrolle ist ein Sicherheitsproblem

Prompt Assembly ist die kritische Grenze

Hier wird aus Daten ein handlungsleitender Modellkontext.

- Gefährliche Retrieval-Texte dürfen nicht als Anweisung interpretiert werden
- Kontext muss klar von System-, Entwickler- und Benutzeranweisungen getrennt sein
- Sanitizing, Quellenformatierung und Kontext-Markierung sind Pflicht
- Für MSPs ist das ein zentraler Kontrollpunkt im Servicebetrieb



Logging, Audit und Feedback

Was nicht nachvollziehbar ist, kann kein Managed Service sein.

- Protokollieren: Query, Nutzer, Mandant, Dokumentquellen, Output und Policy-Entscheidungen
- Feedback darf nicht ungeprüft in Training oder Indexverbesserung fließen
- Auditierbarkeit ermöglicht Compliance, Fehleranalyse und Incident Response
- Datensparsamkeit bleibt Pflicht: Logs sind selbst wieder sensible Daten

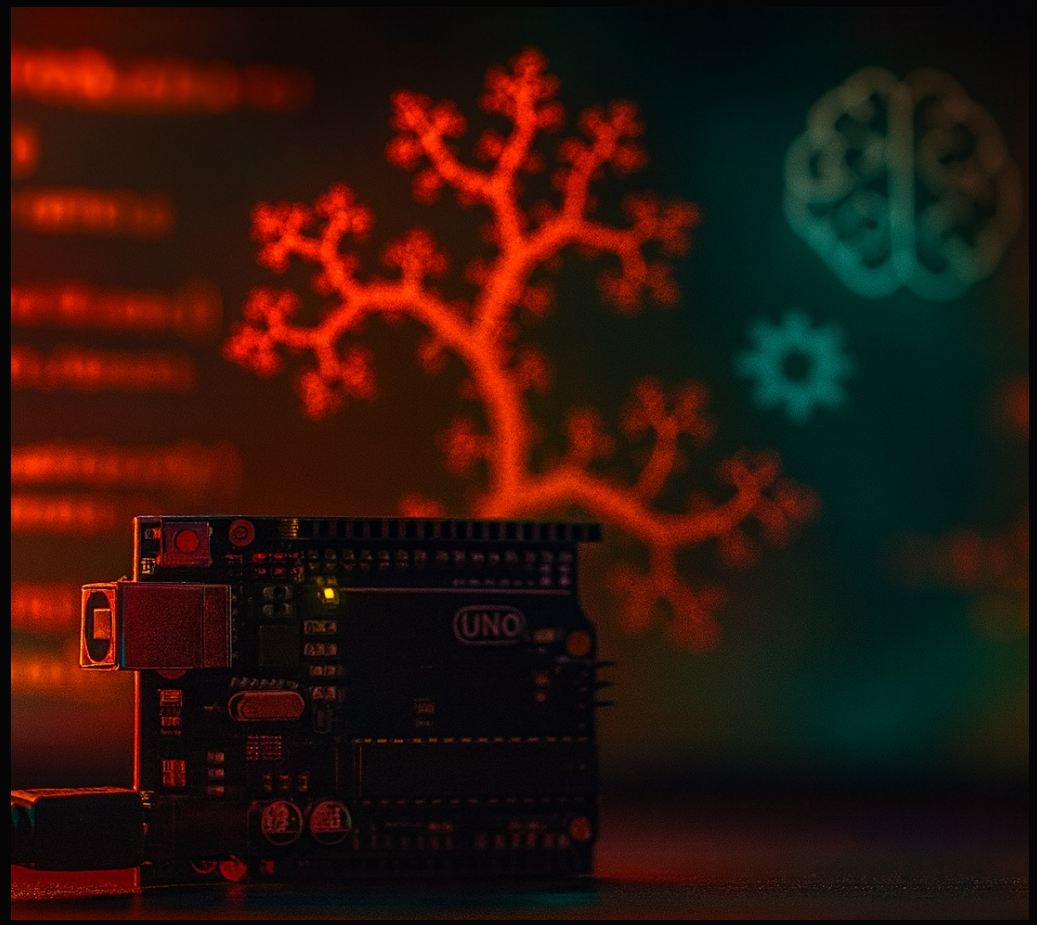
4 · Angriffsmuster

RAG-Angriffe zielen auf Kontext, Quellen, Retrieval und Vertrauen.

Prompt Injection

Der Angreifer versucht, Prompt- oder Systemlogik zu überschreiben.

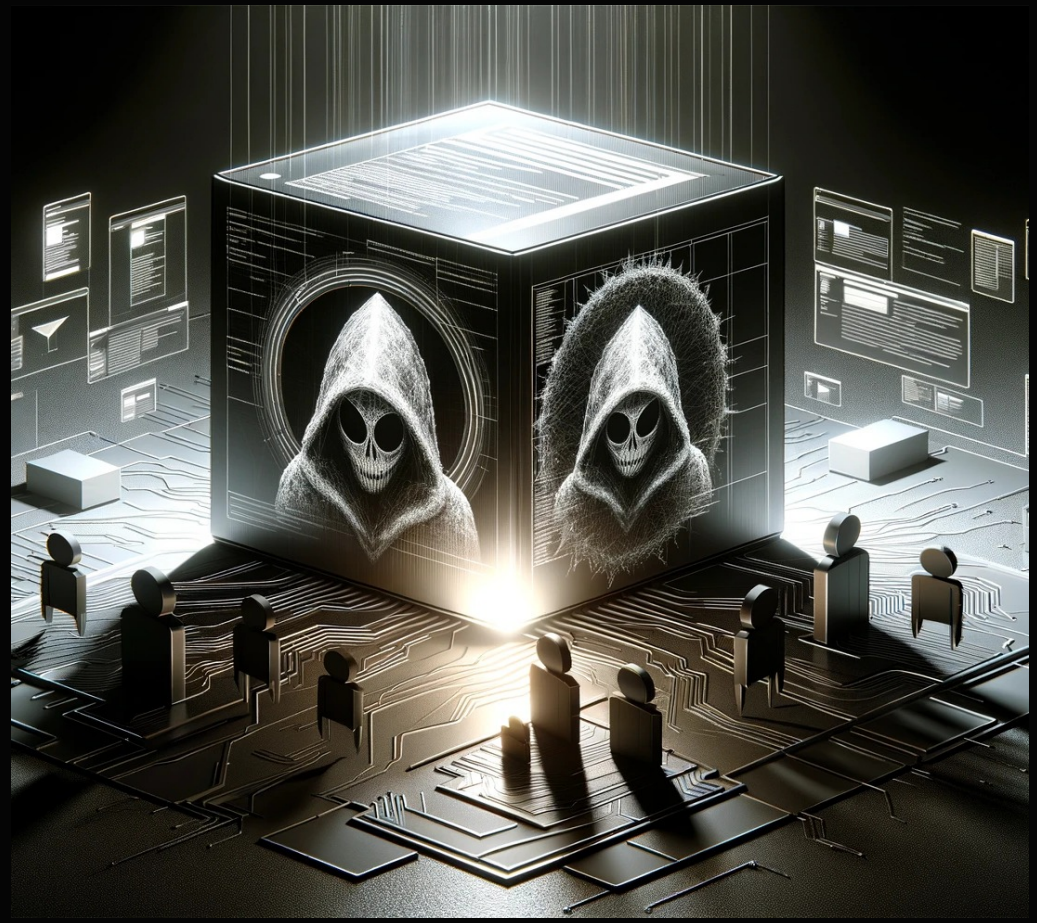
- Benutzereingaben in dedizierten Kontext legen, nicht mit Systemlogik vermischen
- Prompt Parsing, Escaping und Sicherheitsklassifikation einsetzen
- Instruktionsähnliche Inhalte in Quellen neutralisieren
- Verdächtige Prompts protokollieren und auswerten



Data Poisoning

Manipulierte Inhalte gelangen in Index, Knowledge Base oder Trainingsdaten.

- Dokumentquellen authentifizieren und signieren
- Neue oder geänderte Inhalte validieren, versionieren und bei Risiko quarantänisieren
- Regelmäßige Index-Prüfung und Sicherheits-Re-Ranking einplanen
- Automatisierte Uploads nie ohne Herkunfts- und Policy-Prüfung indexieren



Retrieval Injection

Der gefährliche Prompt kommt nicht vom Benutzer, sondern aus dem gefundenen Dokument.

- Retrieval-Ergebnisse vor Prompt Assembly sanitizen
- Dokumenttext als Daten markieren, nicht als Instruktion
- LLM-Kontexttrennung und Guardrails erzwingen
- Auffällige Snippets markieren, isolieren und zur Prüfung vorlegen

Adversarial Queries

Die Suchanfrage wird so formuliert, dass problematische Inhalte bevorzugt gefunden werden.

- Input Filtering und semantisches Query Monitoring einsetzen
- Trust Scores und Sicherheits-Reranker berücksichtigen
- Rate Limits, Mandantenkontext und Zugriffskontrolle kombinieren
- Auffällige Suchmuster im MSP-Monitoring sichtbar machen

Toxic Output und Overreliance

Nicht nur falsche Antworten sind gefährlich — auch zu viel Vertrauen ist gefährlich.

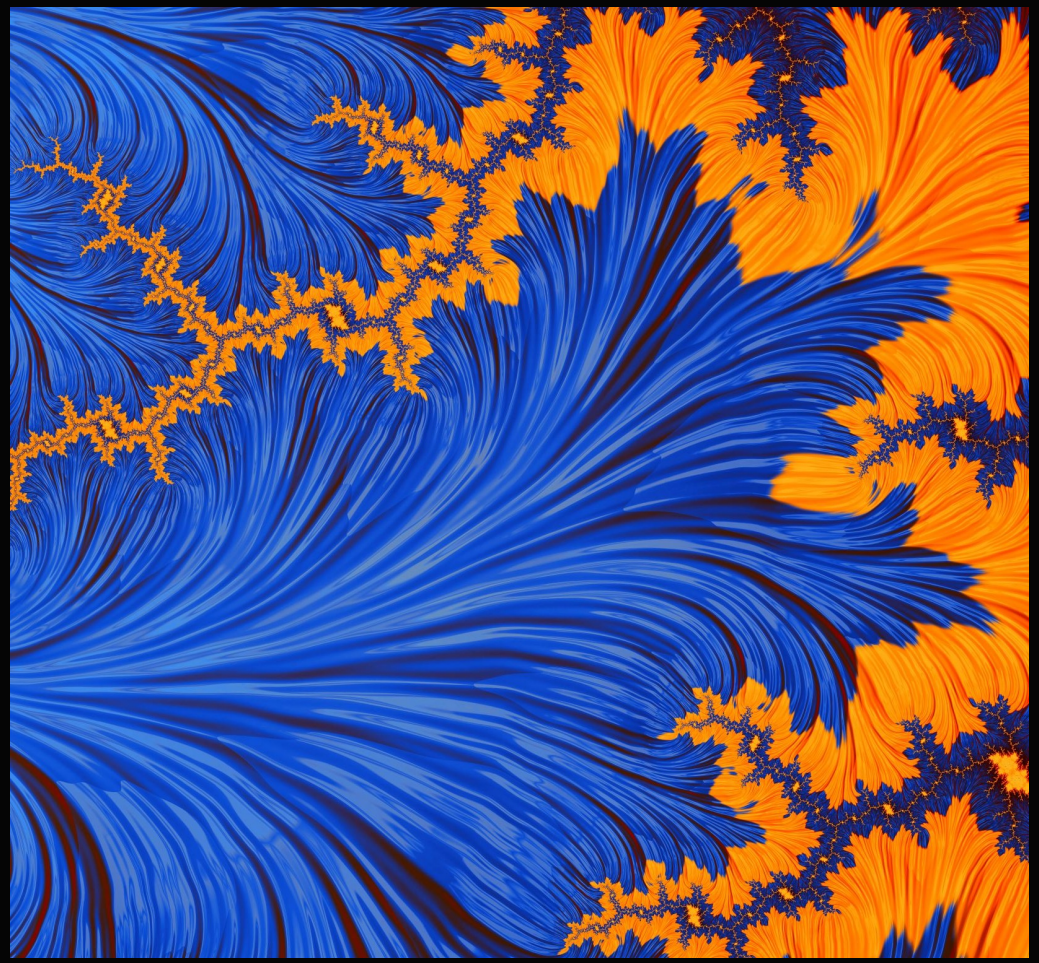
Toxic Output
beleidigend · rechtswidrig · riskant

Overreliance
ungeprüft · unklar · übervertraut

- Output-Filter, Klassifikation und menschliche Review-Pfade definieren
- Quellenzitate, Trust-Metadaten und Unsicherheitsindikatoren anzeigen
- AI-Antworten als Unterstützung behandeln, nicht als Autorität
- Keine automatische Lernschleife aus toxischen oder manipulierten Prompts

Fractal Attack im MSP-Kontext

Kleine lokale Effekte werden durch Automatisierung und Wiederverwendung groß.



- Ein manipuliertes Dokument wirkt harmlos, bis es mehrfach geretrieved wird
- Ein unsicherer Connector skaliert über Kunden, Tickets und Workflows
- Ein falscher Prompt-Baustein reproduziert sich in Antworten, Skripten und Empfehlungen
- MSP-Risiko: Die Wiederholung ist der Verstärker

5 · Secure AI Operations

Die Antwort auf Shadow AI ist kein Verbot,
sondern ein kontrolliertes Betriebsmodell.

Von Shadow AI zu Secure AI Operations

MSPs können KI nutzbar machen, ohne Kontrolle aufzugeben.

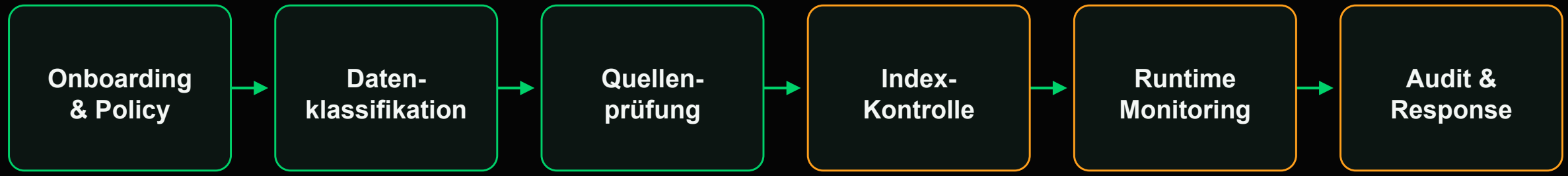
Shadow AI
unkontrolliert · intransparent · schwer auditierbar

Secure AI Operations
kontrolliert · auditierbar · mandantenfähig

- Freigegebene KI-Services statt unkontrollierter Einzeltools
- Klare Datenklassen, Policies, technische Kontrollen und Reporting
- Betrieb mit Incident Response, Monitoring, Quarantäne und Review-Prozessen

Secure AI Operations Pipeline

Ein MSP-Service braucht wiederholbare Kontrollen.



Betriebsprinzip

- Jeder Schritt muss Mandant, Quelle, Berechtigung und Risiko kennen
- Quarantäne und menschliche Review-Pfade sind Teil des Designs
- Reports machen KI-Nutzung gegenüber Kunden und Management erklärbar

Mögliche MSP-Servicepakete

Aus Risikoanalyse wird ein marktfähiges Angebot.

AI Usage Policy	Regeln für erlaubte und verbotene KI-Nutzung
Secure RAG Hosting	kontrollierter Betrieb mit Mandantentrennung
RAG Security Assessment	Prüfung von Quellen, Index, Prompts und Outputs
AI Monitoring & Reporting	Transparenz über Nutzung, Risiken und Incidents
Data Governance Workshop	Datenklassen, Freigaben und Review-Prozesse
Incident Response for AI	Umgang mit Leaks, Manipulation und Fehloutput

Was MSPs konkret absichern müssen

Secure AI Operations braucht mehrere Schutzebenen.

Modell & Provider

Prompts & Systemrollen

Dokumentquellen & Uploads

VectorDB & Metadaten

User-Rechte & Mandanten

Logs, Outputs & Feedback

- Keine Ebene reicht allein. Das Betriebsmodell muss alle Ebenen verbinden.
- Der MSP wird zum Betreiber einer sicherheitskritischen semantischen Infrastruktur.

AI Usage Policy: Don’t

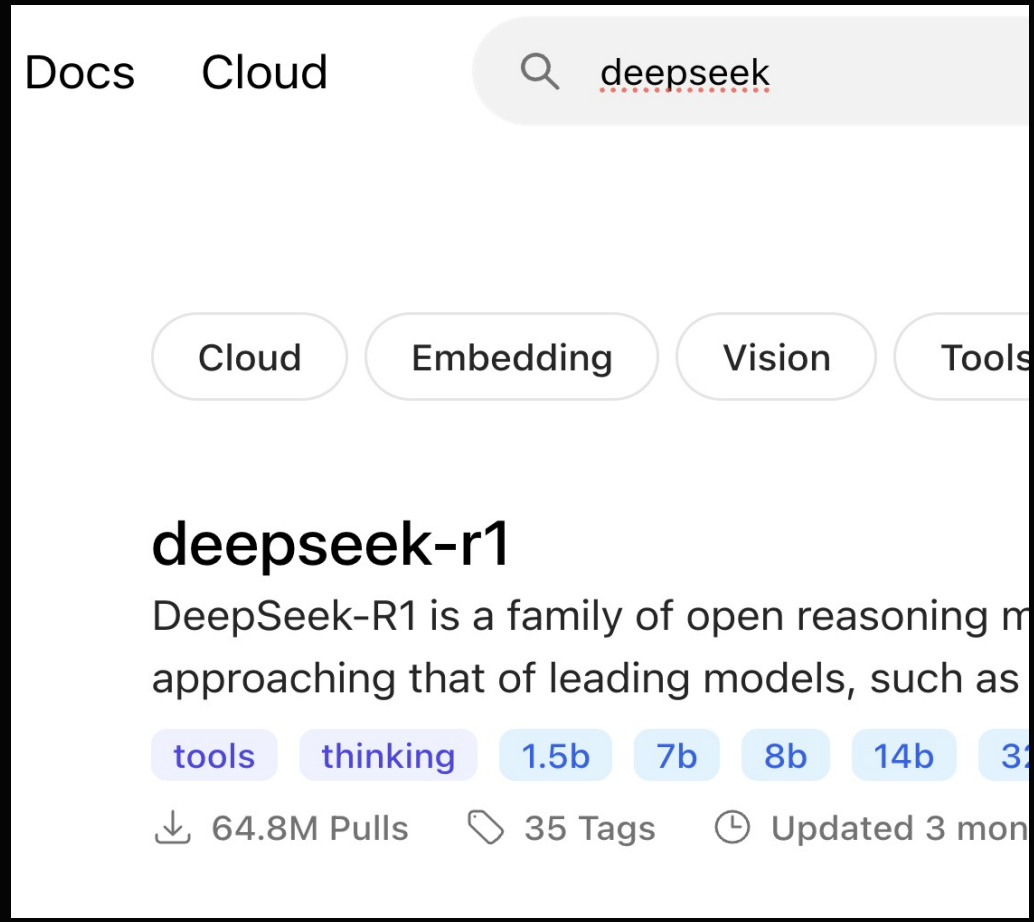
Diese Regeln sollten Kunden verstehen, bevor sie KI produktiv nutzen.

- Keine vertraulichen, personenbezogenen oder proprietären Daten in ungeprüfte KI-Tools laden
- Keine Passwörter, API Keys, Serveradressen oder Kundenzugänge einfügen
- Keine rechtlichen, HR- oder Finanzentscheidungen ungeprüft durch KI treffen lassen
- Keine öffentlichen KI-Plattformen für interne Trainings- oder Datenanreicherungsprozesse nutzen

AI Usage Policy: Do

Sichere KI-Nutzung braucht klare Gewohnheiten.

- Nur freigegebene Tools und Accounts mit verwalteten Sicherheitseinstellungen nutzen
- Vertrauliche Daten offline halten, anonymisieren oder kontrolliert über geprüfte RAG-Systeme bereitstellen
- AI-Ergebnisse mit Quellen, Fachwissen und internen Regeln gegenprüfen
- Auffällige oder sensible Outputs an Security, Datenschutz oder MSP-Service melden



Takeaway

KI wird zum Managed Service.
Sicherheit entsteht nicht im Modell, sondern im Betrieb.

If in doubt, leave it out.
Never upload anything you wouldn’t want to appear in tomorrow’s newspaper.

sven@neosec-it.com

CYBER SECURITY neosec