# Leitfaden: Vorfallreaktion bei KI (Prompt-Leck & Co.)

**In den ersten 60 Minuten die Kontrolle zurückholen**

**Was ist ein KI-Vorfall?**

Ein Sicherheits- oder Datenschutzvorfall, bei dem **Eingaben (Prompts/Eingabetexte), Ausgaben, angebundene Datenquellen (z. B. Dokumentenabruf/RAG) oder KI-Agenten mit Werkzeugzugriff** zu Schaden führen können: Datenabfluss, Manipulation, Fehlentscheidungen, ungewollte Aktionen.

**10-Minuten-Einschätzung (Triage)**

Beantworten Sie drei Fragen – ehrlich, nicht optimistisch:

1. **Sind personenbezogene Daten betroffen?** (Kunden, Mitarbeitende, Protokolle)
2. **Konnte etwas nach außen?** (externer Anbieter, Freigabelink, E-Mail, Screenshot)
3. **Kann es sich wiederholen?** (gleiche Vorlage, gleiche Anbindung, gleiche Berechtigung)

**Wenn 2× „Ja“ oder „Unklar“ → Vorfallmodus aktivieren.**

**Rollen (kleines Kernteam)**

- **Vorfallleitung (IT-Sicherheit):** steuert, entscheidet Eindämmung
- **KI-/Plattform-Verantwortliche:** kennt Modell, Einstellungen, Anbieter, Anbindungen
- **Datenschutz (DSB/DPO):** Risiko, Dokumentation, ggf. Meldung
- **Recht/Kommunikation (bei Bedarf):** externe Aussagen, Kundeninfo

**Sofortplan 60 Minuten (Stabilisieren → Sichern → Überblick)**

**A) Stabilisieren (0–15 Min.)**

- Betroffene KI-Funktion **eingrenzen**: nur intern, nur bestimmte Gruppen oder **temporär pausieren**
- Bei KI-Agenten/Werkzeugen: riskante Aktionen **abschalten**, Berechtigungen **reduzieren**
- Zugangsdaten **rotieren** (Schlüssel/Zugangstoken), falls Missbrauch möglich ist

**B) Beweissicherung (parallel, 0–30 Min.)**
Sichern Sie *vor* dem Aufräumen:

- vollständigen **Eingabetext (Prompt)** + Kontext (z. B. abgerufene Dokumente)
- **Ausgabe** (inkl. Werkzeugaufrufe/„was wurde getan?“)
- relevante **Protokolle** (Anbieter, Proxy, SIEM, Anwendung)

**C) Überblick (30–60 Min.)**

- Was ist der Vorfalltyp? (Prompt-Leck, Datenabfluss, Agent-Aktion, Wissensbasis manipuliert)
- Welche Systeme/Teams sind betroffen?
- Erste Risikoeinschätzung (personenbezogene Daten? Außenwirkung? Wiederholbarkeit?)

**Mini-Impuls (sofort umsetzbar):** Vergeben Sie **eine Vorfall-Kennung** und hängen Sie *alles* daran (Screenshots, Protokolle, Exportdateien, Chatverläufe). Das spart Stunden.

**1) Prompt-Leck (interne Anweisungen / „Systemprompt“ wird sichtbar)**

**Typische Hinweise:** Ausgabe enthält interne Regeln, Tools, Schaltflächen, „versteckte“ Anweisungen.

**Sofort:**

- Vorlagen/Anweisungen **einfrieren** (Version sichern), Freigabelinks **deaktivieren**
- Sensible Inhalte **aus Vorlagen entfernen**, neue Version ausrollen

**Prüfen:**

- Kam das Leck durch **geschickte Eingaben** zustande (Prompt-Manipulation)?
- Stecken Anweisungen „hart“ in **Repos/Wikis** oder Fehlkonfigurationen?

**2) Datenabfluss über Eingaben/Uploads (z. B. PII, Geschäftsgeheimnisse)**

**Typische Hinweise:** Ticket mit Kundendaten im Prompt, Uploads in externe KI, Export in Protokollen.

**Sofort:**

- Betroffene Nutzung **stoppen** (technische Sperre + Hinweis)
- Anbieter-Einstellungen prüfen: **Speicherung/Protokollierung/Training** (Opt-out?)

**Prüfen:**

- Welche Datenkategorien? Wie viele Betroffene ungefähr?
- Wo existieren Kopien? (Anbieter, Browser-Zwischenspeicher, Telemetrie, Backups)

**3) KI-Agent führt ungewollte Aktionen aus (Werkzeug-/Systemzugriff)**

**Typische Hinweise:** unerwartete Änderungen, neue Freigaben, versendete Nachrichten, API-Aufrufe.

**Sofort:**

- Werkzeugzugriffe **begrenzen** (nur lesend, nur genehmigte Aktionen)
- Schlüssel/Zugangstoken **rotieren**, riskante Aktionen **nur mit Bestätigung**

**Prüfen:**

- Kam die Aktion über externe Inhalte (E-Mail/Web/PDF) als **Eingabe-Manipulation**?
- Welche Mindestberechtigungen braucht der Agent wirklich?

**4) Wissensbasis / Dokumentenabruf manipuliert (RAG-Vergiftung)**

**Typische Hinweise:** Antworten plötzlich „überzeugend falsch“, neue Dokumente mit versteckten Anweisungen.

**Sofort:**

- Aufnahme neuer Inhalte **stoppen**, Index auf **Sicherungspunkt** zurücksetzen

**Prüfen:**

- Quellenfreigabe, Signaturen, Quarantäne-Prüfung in der Aufnahme-Pipeline

**Kurzer Datenschutz-Check (für DSB/DPO)**

- **Datenart:** personenbezogen? besondere Kategorien?
- **Außenwirkung:** konnten Daten Dritte erreichen?
- **Risiko:** Identitätsmissbrauch, Diskriminierung, Vertrags-/IP-Schaden?

*(Dokumentation ist Pflicht – auch wenn Sie am Ende nicht melden.)*

**Mini-Vorfallformular (Copy/Paste)**

- Vorfall-Kennung:
- Was ist passiert (1 Satz):
- Betroffene KI-Anwendung/Komponenten:
- Datenkategorien:
- Außenabfluss möglich? (Ja/Nein/Unklar):
- Eindämmung (1 Satz):
- Nächster Schritt + verantwortlich: