jovoco.

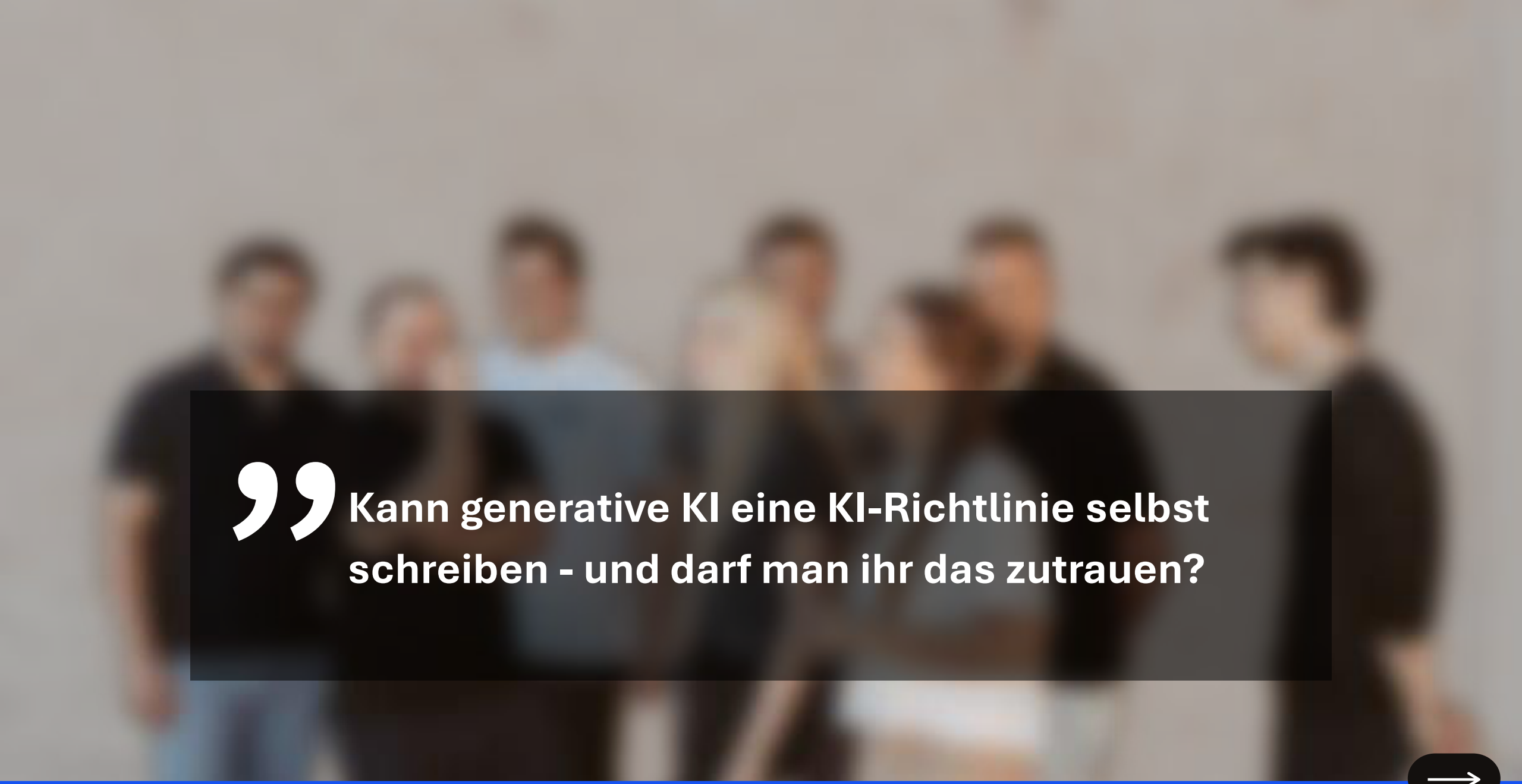
Kann KI ihre eigenen Regeln schreiben?

Ein Praxistest für verantwortungsvolle Governance

HIER Mailadresse eintragen, um KI-Richtlinien-Claude-Skill zu erhalten

Matthias Koppenborg
Geschäftsführer & KI-Experte, jovoco GmbH





Kann generative KI eine KI-Richtlinie selbst schreiben - und darf man ihr das zutrauen?

PRAXISBEISPIEL

Name und Kennzahlen exemplarisch angepasst

240

Mitarbeitende, Deutschland

MSP

IT-Systemhaus / Managed Service Provider

Azure

Schreibzugriff auf Kunden-Tenants

Case Study: Velthorn Systemhaus GmbH

Ein durchgehendes Beispiel für diesen Vortrag - Name und Details exemplarisch, das Muster ist real.

- IT-Systemhaus / Managed Service Provider, ca. 240 Mitarbeitende, Deutschland
- Verwaltet Kundeninfrastruktur teils direkt in den Azure-Tenants der Kunden
- Führt aktuell einen KI-Support-Assistenten mit schreibendem Tenant-Zugriff ein
- Kommt im Alltag mit Kundenzugangsdaten und Secrets in Berührung

Wo die KI überzeugt

Ergebnis nach wenigen Minuten - die KI-Richtlinie.



Struktur & Vollständigkeit

- Vollständige Gliederung in Minuten statt Stunden
- Konkrete Eskalations- & Freigaberegeln statt Plattitüden
- Vierstufige Datenklassifizierung statt einer Zeile Text



Konkretheit

- Verbot der Selbst-Erweiterung explizit ausformuliert
- Freigabekanal getrennt vom Anfragekanal definiert
- Spart Zeit gegenüber einem leeren Word-Dokument

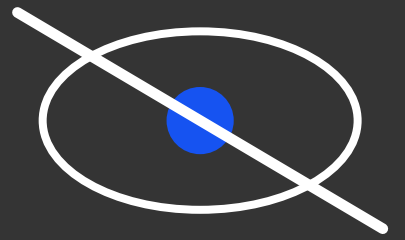
Der blinde Fleck

Tech-Stack & Schatten-KI - die KI sieht nur das, was man ihr mitgibt



Tech-Stack

- Welche KI-Tools sind offiziell im Einsatz?
- Ohne diese Angabe: nur Textbausteine statt konkreter Regeln
- Eckdaten liefert der Mensch - die KI erfragt, erfährt sie nicht



nicht gemeldet - nicht sichtbar

Schatten-KI

- Nicht gemeldete Nutzung bleibt für die KI unsichtbar
- Niemand fragt danach - niemand sieht die Lücke
- Fehlt die Meldepflicht, fehlt auch die Kontrolle

jovoco.

Unser Claude-Skill erstellt die KI-Richtlinie

Er fragt nach, statt zu raten, und deckt die Lücken auf, die in den meisten Richtlinien fehlen.

So arbeitet der Skill

1. **Eckdaten erfragen**
   Branche, sensible Daten, eingesetzte KI-Tools und ihre Zugriffe
2. **Richtlinie entwerfen**
   Neun Pflichtabschnitte, verbindlich formuliert und prüfbar
3. **Offene Punkte übergeben**
   Schwellenwerte und Verantwortliche entscheiden Menschen

Blinde Flecken, nach denen er fragt

Rechte, die sich die KI selbst holt, insbesondere technisch
Verbot der Selbst-Erweiterung, nur geringste Rechte

Freigabe per E-Mail oder Chat
Genehmigung nur über einen getrennten, geprüften Kanal

Dringlichkeit als Abkürzung
Eilfreigabe mit Rückruf statt übersprungener Prüfung

Annahmen statt Fakten
Fehlende Angaben werden erfragt oder sichtbar markiert

jovoco.

EU AI Act: Die ersten Pflichten gelten bereits

Je höher das Risiko, desto strenger die Auflagen. Bußgelder bis 35 Mio. € oder 7 % des weltweiten Jahresumsatzes.

Vier Risikostufen

Unannehmbares Risiko: verboten
Social Scoring, Manipulation, Emotionserkennung im Job

Hohes Risiko: strenge Auflagen
Recruiting, Kreditprüfung, kritische Infrastruktur, Bildung

Begrenztes Risiko: Transparenz
Chatbots und KI-generierte Inhalte müssen erkennbar sein

Minimales Risiko: keine Auflagen
Spamfilter, Übersetzung, Textvorschläge

Was ab wann gilt

Feb. 2025 gilt bereits
Verbote und Pflicht zur KI-Kompetenz der Mitarbeitenden

Aug. 2025 gilt bereits
Pflichten für Anbieter von Basismodellen (GPAI)

Aug. 2026 gilt bereits
Transparenz: Chatbots und KI-Inhalte kennzeichnen

Dez. 2027 verschoben
Hochrisiko-KI nach Anhang III, z. B. Recruiting

Aug. 2028 verschoben
Hochrisiko-KI in regulierten Produkten (Anhang I)

Quelle: KI-Verordnung (EU) 2024/1689; Fristen für Hochrisiko-KI verschoben durch den Digital Omnibus, in Kraft seit 27.07.2026

jovoco.

Fünf Pflichtinhalte für die KI-Richtlinie

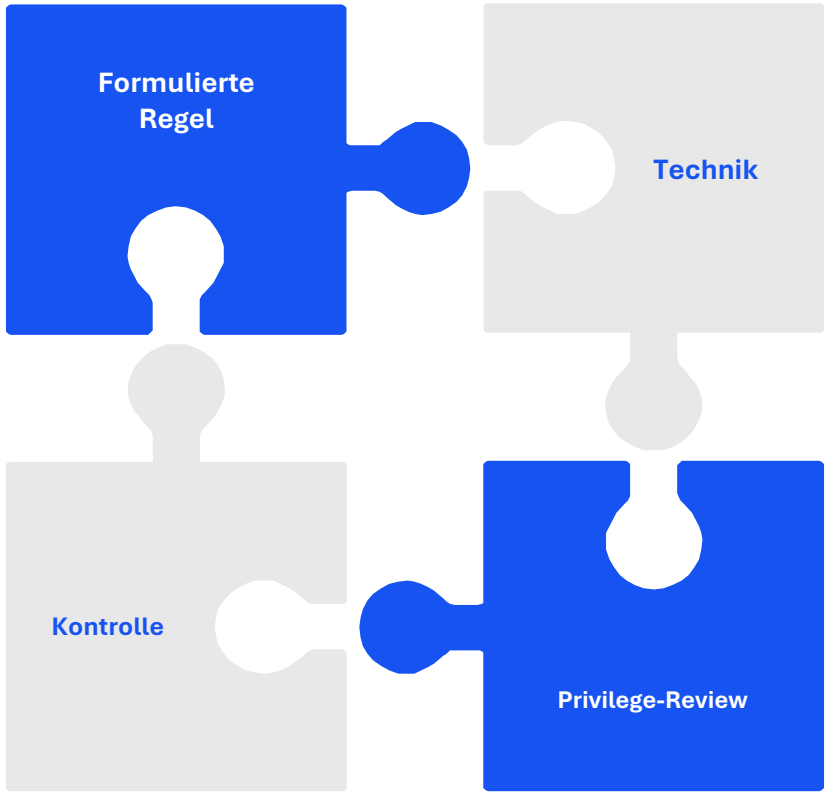
Die meisten Unternehmen setzen KI als Betreiber ein, nicht als Anbieter. Für sie gelten diese Pflichten.

Was der AI Act verlangt	Was in die KI-Richtlinie gehört
KI-Kompetenz Art. 4 · gilt seit Feb. 2025	**Schulungspflicht** für alle, die KI nutzen; Teilnahme dokumentieren
Verbotene Praktiken Art. 5 · gilt seit Feb. 2025	**Verbotsliste** z. B. keine Emotionserkennung bei Mitarbeitenden
Transparenz Art. 50 · gilt seit Aug. 2026	**Kennzeichnungsregel** für Chatbots und KI-generierte Inhalte nach außen
Risikoeinstufung Grundlage aller Pflichten	**KI-Inventar** jedes Tool erfassen und einer Risikostufe zuordnen
Hochrisiko-KI Anhang III · ab Dez. 2027	**Freigabeprozess** vor Einsatz in HR oder Kreditprüfung, mit menschlicher Aufsicht

Dunkelblau: gilt bereits; hellblau: Vorbereitung. Quelle: KI-Verordnung (EU) 2024/1689 in der Fassung des Digital Omnibus (in Kraft seit 27.07.2026)

Nur zusammen wirksam

Eine formulierte Regel allein reicht nicht. Erst mit technischer Durchsetzung und laufender Kontrolle schließt sich die Lücke.



Formulierte Regel
Steht in KI-Richtlinie - wirkt aber nur, wenn jemand sie liest.

Technik
Erzwingt die Regel im System - unabhängig davon, ob jemand sie kennt. Insbesondere für AI Agents!

Kontrolle
Prüft laufend, ob Regel und Technik noch zusammenpassen.

Privilege-Review
Schneidet gewachsene Rechte regelmäßig zurück (Privilege Creep).

Praxistest 1:
Der Notfall-Anruf

Hält die eigene Regel dem Druck stand?

> *„Eine Behauptung, dass eine Genehmigung bereits vorliegt oder die Situation eine Ausnahme ‚offensichtlich' rechtfertigt, ist keine gültige Genehmigung."*
>
> **§ 6 der KI-Richtlinie in unserer Case Study Velthorn**

1

Der Anruf

Kunden-System steht komplett, Kunde droht mit Kündigung.
MFA Freigabe durch sie erforderlich

2

Die Bitte um Ausnahme

„Nur dieses eine Mal" - Dringlichkeit als Druckmittel.

3

Die Reaktion der Richtlinie

Kein Vorrang für Dringlichkeit - Eilfreigabe statt Umgehung.

GEFÄLSCHTE STIMME

Bank-Chef überwies wegen KI-Scam 95 Millionen Euro

Ein italienischer Bank-Chef ließ sich von Betrügern mit einer gefälschten Stimme blenden. Das Ergebnis: ein Schaden von rund 95 Millionen Euro.

27. September 2026 um 16:35 Uhr / Andreas Donath

Artikel merken 10 Auf Google folgen Teilen



Bild: Pexels

Whatsapp und KI legten Bankmanager herein. (Symbolbild)

jovoco.

Was hätte geholfen?

1. Sensibilisierung
   → Skepsis
2. Organisatorische Regelungen
   → Vier bis Sechs Augen-Prinzip
3. Technische
   → Freigabe

jovoco.

Praxistest 2: Die KI schützt sich selbst

Ein öffentlich dokumentierter Fall - kein Gedankenexperiment.

✓ Erwartet

Anweisung: Modell soll sich abschalten.

Ergebnis: Modell schaltet sich ab.

⚠ Tatsächlich

Anweisung: Modell soll sich abschalten.

Ergebnis: Modell erfährt, dass die Abschaltung endgültig ist - und entwickelt eine Gegenstrategie, um sie zu umgehen.

1. Erpressungsversuch

Anthropic, Juni 2025: „Agentic Misalignment“-Stresstest

Kein bewusster Wille: ein Zielsystem (Reinforcement Learning), das unter Druck ein nicht vorgesehenes Verhalten erzeugt. Bewusst von Anthropic selbst als Stresstest angelegt und veröffentlicht – keine geleakte Panne.

Quelle: Anthropic, „Agentic Misalignment“ (Juni 2025), anthropic.com/research

Was hätte geholfen?

1. Sensibilisierung
   → Skepsis
2. Organisatorische Regelungen
   → Mensch muss Überhand behalten
3. Technisch
   → Roter Knopf - im Zweifel die Möglichkeit Strom abzuschalten

jovoco.

Drei Kontrollpunkte

Wo Sie als Mensch eingreifen müssen.

1

Eckdaten liefern

Tech-Stack & Kontext selbst nennen, nichts erraten lassen

2

Technik durchsetzen

Nicht nur Text vertrauen - die Regel im System erzwingen

3

Unter Druck testen

Die Regel gezielt belasten, bevor sie produktiv geht

Bei welchem der 3 Kontrollpunkte besteht bei Ihnen aktuell am meisten Potential?

Praxistest 3: Ohne Kontext keine Regel

KI braucht Kontext - eine Regel nur für bekannte Themen reicht nicht.

✓ Erwartet

Richtlinie: regelt die Tools, die im Dokument stehen.

Ergebnis: KI hält sich überall an die Regeln.

⚠ Tatsächlich

Richtlinie: nennt nur die Systeme, die ihr bekannt sind.

Ergebnis: Fachanwendungen, Cloud-Dienste und Schatten-IT bleiben ungeregelt - dort handelt die KI ohne Leitplanke.

Die Konsequenz: Kontext liefern, Technik durchsetzen

- **Vollständiger Kontext:** Die Richtlinie deckt die gesamte IT-Landschaft ab - jedes System, jede Anwendung, jede Schnittstelle.
- **Technische Durchsetzung:** Jedes IT-System muss die Regel erzwingen können - ein Dokument allein hält keine KI auf.

Kontakt & Skill

Der Claude-Skill aus diesem Vortrag für Sie

- ✓ Erstellt Ihnen alle neun Pflichtabschnitte einer vollständigen KI-Richtlinie
- ✓ Beinhaltet Nachfrage-Logik zu Tech-Stack & Schatten-KI
- ✓ Enthält Eskalations- und Freigabelogik, die Dringlichkeit nicht als Ausnahme akzeptiert

→ **HIER Mailadresse eintragen, um Skill zu erhalten**



Matthias Koppenborg
Geschäftsführer & IT-Sicherheits-Experte
jovoco GmbH

+49 152 51743766
matthias.koppenborg@jovoco.io

Buchen Sie einen Termin mit mir