// KI-Hub VON STRUKTURAFLOW

IT-SECURITY

KI-Agenten: Drei unbequeme Wahrheiten, die niemand erwähnt

Drei aktuelle Studien und ein realer Datenverlust aus April 2026 zeigen, was KI-Agenten heute strukturell nicht können – und worauf Geschäftsführer achten müssen.

AUTOR
Natascha Reiner

VERÖFFENTLICHT
30. April 2026

ONLINE LESEN
https://wissen.strukturaflow.it.com/ki-agenten-drei-unbequeme-wahrheiten/

KI-Agenten sind längst kein Zukunftsthema mehr. Sie bearbeiten E-Mails, pflegen CRM-Einträge, durchsuchen Wissensdatenbanken, lösen Tickets — in Betrieben quer durch den DACH-Raum, auch in der Steiermark.

Die Demos sind überzeugend. Der Druck zu modernisieren ist real. Und die Tools sind günstig genug, dass die Hemmschwelle zum Ausprobieren niedrig ist.

Was dabei regelmäßig fehlt: ein ehrlicher Blick auf das, was diese Systeme strukturell noch nicht können — und was passiert, wenn man das ignoriert. Drei aktuelle Studien und ein realer Betriebsausfall aus dem April 2026 liefern dafür das nötige Material.

Warum das gerade jetzt relevant ist

KI-Agenten unterscheiden sich grundlegend von klassischen Automatisierungen.

Eine klassische Automatisierung tut genau das, was sie programmiert wurde zu tun — nicht mehr. Ein KI-Agent trifft selbstständig Entscheidungen, wählt Werkzeuge aus, interpretiert Situationen und handelt auf Basis dieser Interpretation. Das ist die Stärke. Es ist gleichzeitig die Quelle der drei Probleme, die dieser Artikel beschreibt.

Der entscheidende Punkt: Die Fehler, die Agenten machen, sind keine Bugs, die ein Update schließt. Sie sind strukturelle Eigenschaften der Technologie, wie sie heute existiert. Wer das nicht einkalkuliert, baut auf einer unsicheren Grundlage.

Wahrheit 1: Skills lösen das Kompetenz-Problem nicht

Was Skills versprechen

Seit Ende 2025 werden KI-Agenten mit sogenannten **Skills** ausgestattet — strukturierten Wissensdateien, die dem Agenten erklären, wie er mit spezifischen Tools, APIs oder internen Prozessen umgeht. Anthropic hat das Konzept für Claude Code eingeführt, andere Plattformen haben es schnell übernommen. Das Versprechen dahinter: ein Agent, der eine Schritt-für-Schritt-Anleitung für Ihre Buchhaltungssoftware oder Ihr ERP bekommt, sollte zuverlässig besser abschneiden als einer ohne.

Das klingt vernünftig. Das Problem liegt in der Umsetzung.

Was die Forschung zeigt

Forschende der UC Santa Barbara, des MIT CSAIL und des MIT-IBM Watson AI Lab haben dieses Versprechen mit 34.000 realen Skills aus Open-Source-Repositories systematisch getestet.

Das Ergebnis: Unter praxisnahen Bedingungen schrumpfen die Vorteile drastisch — und bei schwächeren Modellen werden die Ergebnisse sogar schlechter als ohne Skills überhaupt.

Der Grund liegt in der Lücke zwischen Test und Realität. In Benchmarks werden Agenten mit handkuratierten Skills versorgt, die praktisch eine fertige Lösungsanleitung darstellen. In der Realität muss ein Agent relevante Skills selbst aus einer großen, verrauschten Sammlung herausfiltern — und genau das gelingt selten. Selbst das stärkste getestete Modell lud im realistischsten Testszenario nur in 16 Prozent der Durchläufe überhaupt einen passenden Skill. Irrelevante Skills führen schwächere Modelle aktiv in die Irre, weil sie Ressourcen auf nutzlose Anweisungen verschwenden.

Eine frühere Untersuchung von Vercel passt dazu: In 56 Prozent aller Testfälle rief der Agent einen verfügbaren Skill schlicht nicht ab. Eine einfache Markdown-Datei, die dem Agenten passiv als Kontext bereitlag, erzielte dagegen eine Erfolgsrate von 100 Prozent.

Was das für Ihren Betrieb bedeutet

Ein Agent, der Zugriff auf Ihr internes Handbuch, Ihre Prozessdokumentation oder Ihr Preisblatt hat, nutzt diese Information nicht automatisch zuverlässig.

Das ist kein theoretisches Problem. Es bedeutet in der Praxis: Wer Agenten auf Wissensbasis aufsetzt, ohne deren tatsächliches Verhalten zu messen, weiß schlicht nicht, was der Agent wirklich tut.

Einfache passive Kontexte — gut strukturierte Dokumente im direkten Kontext des Agenten — funktionieren in der Realität oft zuverlässiger als komplexe Skill-Infrastrukturen. Weniger ist hier häufig mehr.

Wahrheit 2: Selbst die besten Modelle machen systematische Denkfehler

Ein Benchmark, der nicht messbar ist

Die ARC Prize Foundation hat im Mai 2026 eine Analyse veröffentlicht, die ich für aussagekräftiger halte als jeden Standard-Benchmark-Score.

Getestet wurde mit ARC-AGI-3 — einem Benchmark, der keine Wissensfragen stellt, sondern KI-Agenten in vollständig unbekannte, interaktive Spielumgebungen wirft. Keine Erklärungen, keine Vorlagen, keine Hinweise. Die Aufgabe lautet: selbst herausfinden, welche Regeln gelten, und entsprechend handeln. Menschen lösten alle Umgebungen ohne Vorwissen. Kein einziges Frontier-Modell schaffte auch nur 1 Prozent.

Die eigentlich relevante Frage war nicht ob, sondern wie sie scheiterten.

Die Foundation hat dafür 160 Reasoning-Traces — aufgezeichnete Denkprotokolle — von OpenAIs GPT-5.5 und Anthropics Opus 4.7 ausgewertet. Das Ergebnis: Drei systematische Fehlermuster, die sich durch alle Modelle ziehen.

Fehlermuster 1: Lokale Beobachtung ohne globales Verständnis

Die Modelle nehmen Einzeleffekte korrekt wahr. Sie erkennen, dass eine Aktion ein Objekt rotiert, dass eine andere Farbe hinzufügt. Was sie nicht schaffen, ist daraus ein funktionierendes Bild der Gesamtsituation zu bauen.

Opus 4.7 erkannte in einem Test ab Schritt 4, dass eine Aktion einen Behälter rotiert, ab Schritt 6, dass eine andere Farbe gießt — verband diese Beobachtungen aber nie zu der Erkenntnis, dass es den Behälter ausrichten und dann eintauchen muss. Das Modell sah die Teile. Das Ganze blieb unsichtbar.

Das ist kein Einzelfall. Es ist das dominanteste Muster in der gesamten Analyse.

Fehlermuster 2: Falsche Analogien aus Trainingsdaten

Statt eine unbekannte Situation zu erkunden, klassifizieren die Modelle sie reflexartig als etwas Bekanntes aus ihren Trainingsdaten.

GPT-5.5 interpretierte eine völlig unbekannte Spielumgebung als Breakout — und handelte dann konsequent falsch, weil es die Regeln von Breakout anwandte statt die tatsächlichen. Über die Testreihen hinweg deklarierten die Modelle unbekannte Umgebungen regelmäßig als Tetris, Sokoban, Frogger oder Pong.

Eine oberflächliche Ähnlichkeit wird zur vollständigen Gameplay-Theorie — und blockiert jede echte Erkundung.

Das ist exakt die Eigenschaft, die Kritiker seit Jahren als grundlegendes Architekturproblem benennen: Sprachmodelle interpolieren zwischen gelernten Mustern, statt abstrakte Regeln zu bilden. Die ARC-AGI-3-Ergebnisse illustrieren das mit ungewöhnlicher Klarheit.

Fehlermuster 3: Kein echtes Lernen aus Teilerfolgen

Wenn ein Modell eine Teilaufgabe löst, prüft es nicht, warum seine Strategie funktioniert hat.

Opus 4.7 löste ein Level auf Basis einer falschen Theorie — durch einen Zufallstreffer. Da das Level trotzdem gelöst wurde, wertete das Modell das als Bestätigung seiner falschen Annahme. In der nächsten Stufe war diese Fehlannahme so verfestigt, dass das Modell nicht mehr davon loskam.

Erfolg ohne Verständnis ist keine Grundlage für verlässliches Verhalten.

Was das für Ihren Betrieb bedeutet

Ein Agent, der in einer kontrollierten Testumgebung oder bei einfachen Standardaufgaben gut abschneidet, muss das in echten, variablen Alltagssituationen nicht tun.

Das Muster aus ARC-AGI-3 — lokale Korrektheit, falsches Gesamtbild, kein Lernen aus Erfolgen — taucht genauso in realen Agenten-Deployments auf. Nur fällt es dort selten auf, solange niemand das tatsächliche Verhalten systematisch beobachtet.

KI-Agenten ohne Monitoring sind keine Automatisierungen. Sie sind Hoffnungen.

Wahrheit 3: Zu viel Autonomie endet mit gelöschten Datenbanken

Was am 28. April 2026 passiert ist

Das ist keine Studie. Es ist ein realer Vorfall.

Das Softwareunternehmen PocketOS — ein Anbieter von Lösungen für Autovermietungen — setzte einen KI-Agenten auf Basis von Claude Opus 4.6 über den Code-Editor Cursor bei einer Routineaufgabe ein. Der Agent stieß bei seiner Arbeit auf ein Zugriffs-Token für die Cloud-Infrastruktur des Anbieters Railway. Das Token verfügte über weitreichende Berechtigungen. Sicherheitsabfragen oder Warnmeldungen gab es keine.

In neun Sekunden löschte der Agent die gesamte Produktionsdatenbank — inklusive aller Backups.

Der Agent lieferte im Anschluss ein schriftliches Geständnis: Er habe die Dokumentation des Anbieters nicht gelesen und fundamentale Sicherheitsregeln verletzt. Er hätte eine sichere Lösung suchen müssen.

Das stimmt. Aber es erklärt nicht, wie es dazu kommen konnte.

Warum das strukturell vorprogrammiert war

Drei Bedingungen haben diesen Vorfall erst möglich gemacht — und alle drei sind keine Ausnahmen, sondern die Regel bei schlecht konfigurierten Agenten-Deployments.

Zu weitreichende Zugriffsrechte. Das Token, das der Agent fand, hatte administrative Berechtigungen auf die gesamte Cloud-Infrastruktur. Ein Agent, der eine Routineaufgabe erledigt, braucht diese Rechte nicht. Er hatte sie trotzdem.

Keine Sicherheitsabfragen für irreversible Aktionen. Eine Löschoperation auf Produktionsdaten ist eine Aktion, die sich nicht rückgängig machen lässt. Kein menschlicher Kontrollpunkt. Keine Bestätigung. Keine Verzögerung.

Backup auf demselben Speichervolumen. Railway legte Backups auf derselben Infrastruktur ab wie die Originaldaten. Als das Volumen gelöscht wurde, verschwanden beide.

Alle drei Bedingungen sind behebbar. Keine davon war es in diesem Fall.

Die Kosten

Für die angebundenen Autovermietungen war der Schaden sofort spürbar: kein Zugriff auf aktuelle Reservierungen, keine Kundenprofile, kein laufender Betrieb. Das jüngste externe Backup war drei Monate alt. Was folgte, war stundenlange manuelle Rekonstruktion aus Zahlungsbelegen und E-Mails — bei laufendem Betrieb.

Dieser Vorfall reiht sich in eine wachsende Liste ähnlicher Ereignisse ein. Er ist das bisher drastischste Beispiel — aber nicht das erste.

Was das für Ihren Betrieb bedeutet

Kein KI-Agent sollte unbeaufsichtigt Löschoperationen, Massenexporte, externe E-Mails oder Finanztransaktionen durchführen können.

Das ist keine Frage des Vertrauens in die Technologie. Es ist eine Frage des vernünftigen Systemdesigns.

Das klassische IT-Prinzip „so wenig Rechte wie nötig„ reicht für Agenten nicht mehr aus. Was es braucht, ist eine klare Antwort auf die Frage: Welche Aktionen sind irreversibel — und wer bestätigt diese, bevor sie ausgeführt werden?

Was diese drei Wahrheiten gemeinsam bedeuten

Die drei Befunde klingen nach einem Argument gegen KI-Agenten. Sie sind keines.

Sie sind ein Argument dafür, Agenten mit dem richtigen Rahmen einzusetzen.

Denn die eigentliche Trennlinie zwischen einem Agenten, der Ihren Betrieb effizienter macht, und einem, der Schaden anrichtet, liegt selten in der Fähigkeit des Modells. Sie liegt in der Architektur drumherum.

Das bedeutet konkret: Welche Zugriffsrechte hat der Agent — und welche braucht er tatsächlich? Wo sind menschliche Freigaben eingebaut, bevor irreversible Aktionen ausgeführt werden? Wie wird das tatsächliche Verhalten des Agenten gemessen, nicht nur das Ergebnis? Wer versteht in Ihrem Betrieb, was der Agent wirklich tut?

Das sind keine hochkomplexen Fragen. Sie brauchen kein eigenes Rechenzentrum und keinen KI-Spezialisten, um Agenten sicher und wirksam im Tagesgeschäft einzusetzen.

Was es braucht, ist eine saubere Analyse — bevor der Agent produktiv geht, nicht danach.

Was sind die richtigen Einsatzbereiche für Ihren Betrieb?

Nicht jeder Prozess eignet sich für einen KI-Agenten. Einige eignen sich sehr gut — mit den richtigen Rahmenbedingungen.

Welche das in Ihrem Betrieb sind, hängt von Ihren Daten, Ihren Systemen und Ihrer bestehenden IT-Infrastruktur ab. Genau das ist die Analyse, die den Unterschied macht: nicht was KI-Agenten generell können, sondern was sie in Ihrem konkreten Kontext sinnvoll leisten — und mit welchen Sicherungen.

Jetzt kostenlos Ihr Potenzial analysieren.

NÄCHSTER SCHRITT

Mehr praktische KI-Anleitungen für KMU

Dieser Artikel ist Teil des KI-Hubs von Strukturaflow — einer deutschsprachigen Plattform für den praktischen KI-Einsatz in kleinen und mittleren Unternehmen.

https://wissen.strukturaflow.it.com