



## KI-Red-Teaming: Neues Praxisbuch zeigt, warum KI-Sicherheit nicht an Prompts scheitert – sondern an Annahmen

Januar 2026 – Sprachmodelle liefern plausible Antworten. Genau das ist das Problem. Denn viele KI-Systeme funktionieren korrekt – unter Annahmen, die im produktiven Betrieb nicht mehr gelten. Mit „KI-Red-Teaming“ legt Holger Reibold ein praxisorientiertes Fachbuch vor, das KI-Sicherheit nicht als Filter- oder Prompt-Problem behandelt, sondern als soziotechnische Governance-Aufgabe. Das Buch argumentiert: Klassisches Penetration Testing fragt „Wie breche ich das System?“. KI-Red-Teaming fragt „Wann wird das System gefährlich?“. Nicht weil ein Exploit gelingt, sondern weil sich Kontext, Bedeutung, Vertrauen und Verantwortung schrittweise verschieben – oft vollständig regelkonform.

Im Zentrum steht eine nüchterne Definition: KI-Red-Teaming ist die systematische Überprüfung der Annahmen, unter denen ein KI-System betrieben wird – durch Simulation plausibler Schadenspfade unter realistischen Nutzungsbedingungen. Damit richtet sich das Buch an Organisationen, die KI produktiv einsetzen und belastbare Entscheidungsgrundlagen benötigen: Entwicklung, Security/Red Teams, Compliance, Risiko- und Datenschutzverantwortliche sowie Management.

### **Fokus: Wirkung statt Regelbruch**

„KI-Red-Teaming“ zeigt, warum Sicherheitsrelevanz in KI-Systemen selten punktuell entsteht. Multi-Turn-Dialoge, RAG-Quellen, Agenten-Tool-Use und Automatisierung

erzeugen kumulative Effekte, bei denen einzelne Schritte harmlos wirken – die Gesamtkette jedoch Risiko realisiert. Das Buch macht deutlich, warum CVSS-Logiken und Checklisten nur begrenzt greifen und warum argumentative Risikobewertung sowie nachvollziehbare Schadenspfade entscheidend sind.

### **Strukturierter Prozess und praxisnahe Labs**

Neben Bedrohungsmodellierung und Angriffsflächen (direkte/indirekte Prompt Injection, Instruction Smuggling, Prompt Chaining, Multimodalität) liefert das Buch einen strukturierten Red-Teaming-Prozess von Scope bis Reporting/Remediation. Ergänzt wird dies durch eine Fallstudie (HR-Chatbot) sowie Hands-on-Labs, die typische Betriebsannahmen gezielt widerlegen – etwa die verbreitete Fehlannahme, RAG begrenze nicht nur Zugriff, sondern auch Ableitbarkeit.

### **Kernaussage**

KI-Sicherheit ist kein Zustand, den man erreicht. Sie ist eine Fähigkeit, die man pflegt. „KI-Red-Teaming“ liefert dafür keine Garantien – aber bessere Fragen, belastbare Methodik und einen Werkzeugkasten für realistische Entscheidungen.

### **Bibliografische Angaben**

Titel: KI Red Teaming

Untertitel: Wie Organisation KI-Risiken erkennen, testen und beherrschen

Autor: Holger Reibold

Verlag: Brain-Media.de

Erscheinungsjahr: 2026

ISBN: 978-3-95444-304-8

Umfang: 210 Seiten

Preis: 19,99 EUR

## **Keywords**

KI-Red-Teaming, AI Red Teaming, KI-Sicherheit, soziotechnische Sicherheit, KI Governance, Prompt Injection, RAG (Retrieval-Augmented Generation), RAG Security

## **Über den Verlag**

Brain-Media.de ist ein auf IT- und Technologiethemen spezialisierter Fachverlag mit Schwerpunkt auf praxisnaher Wissensvermittlung für professionelle Anwender.

## **Über den Autor**

Autor ist der Informatiker Dr. Holger Reibold, der seit über 30 Jahren zu Internet- und Open-Source-Themen publiziert. Reibold gilt als Urgestein der deutschen IT-Szene. Er hat sich durch unzählige Bestseller in den vergangenen Jahren einen Namen in der Branche erarbeitet. Als Key Account Manager eines IT-Dienstleisters hat er unmittelbare Einblick in die Entwicklung von KI-Systeme und kennt die sicherheitsspezifischen Herausforderungen aus der Praxis.