Holger Reibold

KI Incident Response

Wie man Sicherheitsvorfälle in KI-Systemen erkennt, eindämmt und verantwortet

BRAIN-MEDIA.DE

Holger Reibold

KI Incident Responce

Wie man Sicherheitsvorfälle in KI-Systemen erkennt, eindämmt und beherrscht

BRAIN-MEDIA.DE

ISBN: 978-3-95444-306-2

Cover: Freepik

Druck: Libri Plueros GmbH, Friedensallee 273, 22763 Hamburg

Brain-Media.de – St. Johanner Str. 41-43 – 66111 Saarbrücken

info@brain-media.de – www.brain-media.de

Inhaltsverzeichnis

Prolog

KI-Sicherheitsvorfälle sind kein Randphänomen. Sie sind kein Zukunftsthema, kein Forschungsproblem und kein „Edge Case“ für besonders innovative Organisationen. Sie sind Realität – heute, in produktiven Systemen, mit realen Auswirkungen auf Menschen, Märkte und Vertrauen. Vielfach wird KI Incident Response primär als operative Disziplin bewertet: Erkennen, Analysieren, Eindämmen und Beheben von Vorfällen in KI-Systemen. Doch neuere Literatur machen deutlich, dass dieser Blick zu kurz greift. KI Incident Response entwickelt sich zu einem integralen Bestandteil von KI-Risikomanagement, Post-Market-Governance und Wertschöpfungsketten-Verantwortung. Fünf Einsichten sind dabei zentral.

1. KI-Incidents sind erwartbar – nicht außergewöhnlich

Empirische Erhebungen und Incident-Datenbanken zeigen, dass KI-Vorfälle regelmäßig auftreten, oft mit wiederkehrenden Mustern, aber in sehr unterschiedlichen technischen und organisatorischen Kontexten. Die Arbeit des Center for Security and Emerging Technology macht deutlich, dass Incidents nicht als binäre Ereignisse zu verstehen sind, sondern als Spektrum von Near Misses, Fehlverhalten und manifestem Schaden, das systematisch erfasst und ausgewertet werden muss.

Damit verschiebt sich der Fokus: Weg von der Frage, ob ein KI-Incident eintritt, hin zur Frage, wie früh er erkannt wird und wie strukturiert reagiert werden kann. Incident Response ist damit keine Ausnahmesituation mehr, sondern ein kontinuierlicher Feedback-Mechanismus beim Betrieb von KI-Systemen.

2. Incident Response ist Teil von KI-Risikomanagement – nicht dessen Nachsatz

Sowohl das NIST AI Risk Management Framework als auch das zugehörige AI RMF Playbook verorten Incident Response klar innerhalb eines zyklischen Modells aus Govern, Map, Measure und Manage. Reaktion auf Incidents ist dort kein isolierter Prozess, sondern ein Mechanismus mit folgenden Zielen: Risikoeinschätzung aktualisieren, Kontrollwirksamkeit überprüfen und organisatorische Verantwortlichkeiten testen.

Gleichzeitig zeigt die Überarbeitung der klassischen Incident-Response-Leitlinien in NIST SP 800-61r3, dass Incident Response zunehmend als Teil des übergeordneten Risikomanagements verstanden wird – mit klarer Verbindung zu Governance, Dokumentation und Entscheidungsprozessen. Für KI-Systeme bedeutet das konkret: Incident Response erzeugt neue Risikoinformation, diese Information muss zurück in Design, Deployment und Governance gespiegelt werden, andernfalls bleibt sie wirkungslos.

3. Post-Market-Korrekturen werden zur Schlüsselkompetenz

Ein besonders wichtiger Beitrag der neueren Forschung ist der Perspektivwechsel von reiner Reaktion hin zu gezielten Deployment-Korrekturen. Das Framework zu Deployment Corrections for Frontier AI Models beschreibt Incident Response als Fähigkeit, laufende Systeme kontrolliert zu verändern, ohne sie zwangsläufig vollständig abzuschalten. Diese Korrekturen reichen von Nutzer- und Zugriffsrestriktionen, Funktions- und Fähigkeitsbegrenzungen, Output-Filterung, bis hin zur vollständigen Deaktivierung. Entscheidend ist aber: Diese Maßnahmen müssen vorab technisch und organisatorisch vorbereitet sein. Incident Response ohne vorbereitete Korrekturmechanismen reduziert sich auf Improvisation.

4. Wertschöpfungsketten schlagen Systemgrenzen

Die Analyse zur technologischen Neutralität des EU AI Acts macht deutlich, dass KI-Incidents selten entlang klarer Anbieter- oder Rollenmodelle verlaufen. Modelle, Systeme, Datenquellen, Plattformen und Deployments bilden netzwerkartige Wertschöpfungsketten, in denen Informationen über Incidents geteilt werden müssen, um wirksam reagieren zu können. Damit wird Incident Response zu einer kooperativen Aufgabe über Organisationsgrenzen hinweg zwischen Modellanbietern und Systemintegratoren, zwischen Deployern und Plattformbetreibern sowie zwischen technischen und regulatorischen Akteuren. Rigide definierte Zuständigkeiten helfen hier wenig. Gefordert ist wertschöpfungs-

ketten-neutrale Incident-Kommunikation, wie sie auch von OECD und GPAI mit Blick auf gemeinsame Reporting-Formate gefordert wird.

5. Incident Response wird prüf- und berichtspflichtig

Mit dem EU AI Act, begleitenden Konformitätsverfahren wie capAI und sektoralen Taxonomien (z. B. im Gesundheitsbereich) wird Incident Response zunehmend auditierbar. Organisationen müssen künftig nachweisen können, dass sie Incidents erkennen können, dass sie geeignete Reaktionsmaßnahmen definiert haben und dass sie aus Vorfällen systematisch lernen. Incident Response erzeugt damit regulatorisch relevante Artefakte: Incident Reports, Root-Cause-Analysen, Korrekturentscheidungen und Governance-Anpassungen KI Incident Response ist heute weder rein technisch noch rein regulatorisch. Sie ist eine sozio-technische Betriebskompetenz, die Engineering, Security, Governance und Recht zusammenführt. Nicht die Abwesenheit von KI-Incidents ist das Maß für Reife. Reife zeigt sich darin, dass Incidents früh erkannt, klar klassifiziert, kontrolliert korrigiert und systematisch rückgekoppelt werden. In diesem Sinne ist KI Incident Response kein Zeichen des Scheiterns von KI-Systemen, sondern ein Indikator dafür, dass sie unter Kontrolle betrieben werden.

Ich wünsche Ihnen dabei viel Erfolg.

Holger Reibold

Vorwort

Künstliche Intelligenz ist in den operativen Kern vieler Organisationen vorgedrungen. KI-Systeme treffen Entscheidungen, priorisieren Informationen, steuern Prozesse und interagieren zunehmend autonom mit Menschen, Daten und anderen Systemen. Damit wächst nicht nur ihr Nutzen, sondern auch die Zahl und Tragweite von Situationen, in denen KI-Systeme unerwartet, fehlerhaft oder schädlich agieren.

Solche Situationen sind keine hypothetischen Zukunftsszenarien. Sie treten heute auf – in produktiven Systemen, unter realen Einsatzbedingungen, oft außerhalb der Annahmen, unter denen diese Systeme entworfen wurden. Dennoch fehlt es in vielen Organisationen an klaren Antworten auf grundlegende Fragen:

Wann liegt ein KI-Incident vor – und wann nicht?

Wer ist verantwortlich, wenn Ursachen über Daten, Modelle, Deployment und Nutzung verteilt sind?

Welche technischen und organisatorischen Maßnahmen sind im Ernstfall sinnvoll?

Wie lassen sich Vorfälle dokumentieren, melden und systematisch auswerten?

Dieses Buch ist aus der Beobachtung entstanden, dass klassische Incident-Response-Ansätze diese Fragen nur unzureichend beantworten.

Sie setzen deterministische Systeme, klare Systemgrenzen und reproduzierbare Fehler voraus – Annahmen, die auf moderne KI-Systeme oft nicht zutreffen. Gleichzeitig zeigt sich, dass rein ethische, rechtliche oder sicherheitspolitische Debatten wenig helfen, wenn operative Entscheidungen unter Zeitdruck getroffen werden müssen.

Ziel dieses Buches ist es daher, KI Incident Response als eigenständige, aber anschlussfähige Disziplin zu beschreiben; technisch fundiert, organisatorisch umsetzbar und regulatorisch einbettbar. Dabei versteht sich dieses Werk nicht als Sammlung von Worst-Case-Szenarien oder als Warnschrift, sondern als Arbeitsgrundlage für den professionellen Betrieb von KI-Systemen. Incident Response wird hier nicht als Ausnahmezustand verstanden, sondern als notwendige Fähigkeit in einer Umgebung, in der Fehlverhalten, Missbrauch und unerwartete Effekte systemimmanent sind. Das Buch richtet sich an all, die KI nicht nur entwickeln oder einsetzen, sondern für ihren Betrieb Verantwortung tragen.

Das Buch setzt kein tiefes mathematisches oder modelltheoretisches Vorwissen voraus. Technische Konzepte werden so weit erläutert, wie sie für das Verständnis von Incidents und Reaktionen notwendig sind. Code, Algorithmen oder Modellarchitekturen stehen bewusst nicht im Vordergrund – entscheidend ist ihr Verhalten im Betrieb.

Dieses Buch ist kein Handbuch für das „perfekte“ KI-System, keine vollständige Sicherheitsarchitektur und keine Anleitung zur Vermeidung aller Risiken. Es geht stattdessen um den Umgang mit dem Unvermeidlichen: Situationen, in denen KI-Systeme anders agieren als erwartet – und Organisationen dennoch handlungsfähig bleiben müssen.

1 Was ist ein KI-Incident?

Künstliche Intelligenz wird zunehmend in Systemen eingesetzt, deren Fehlverhalten nicht nur technische, sondern auch rechtliche, wirtschaftliche und gesellschaftliche Konsequenzen haben kann. Dennoch fehlt es bislang an einem einheitlichen Verständnis dafür, wann ein solches Fehlverhalten als sicherheits- oder governance-relevanter Vorfall zu behandeln ist. In der Praxis werden unter dem Begriff „KI-Incident" sehr unterschiedliche Sachverhalte zusammengefasst – von gewöhnlichen Modellfehlern über gezielte Angriffe bis hin zu regulatorischen Verstößen ohne unmittelbaren technischen Defekt. Diese begriffliche Unschärfe erschwert nicht nur die operative Reaktion, sondern unterminiert auch die systematische Vorbereitung auf Vorfälle.

Kapitel 1 schafft die begriffliche Grundlage für das gesamte Buch. Es klärt, was im Kontext von KI-Systemen sinnvollerweise als Incident verstanden werden kann, wie sich Incidents von Fehlern, Risiken und Schäden abgrenzen lassen und warum klassische Incident-Definitionen für KI nur eingeschränkt geeignet sind. Damit legt das Kapitel den Rahmen für alle weiteren Überlegungen zur Erkennung, Analyse und Behandlung von KI-bezogenen Vorfällen und macht deutlich, dass Incident Response bei KI-Systemen primär eine Frage des beobachtbaren Systemverhaltens und des daraus resultierenden Handlungsbedarfs ist.

1.1 Warum eine präzise Definition notwendig ist

Incident Response setzt begriffliche Klarheit voraus. In klassischen IT- und Softwaresystemen ist diese Klarheit historisch gewachsen: Ein Incident bezeichnet dort ein unerwünschtes Ereignis, das die Verfügbarkeit, Integrität oder Vertraulichkeit eines Systems beeinträchtigt und eine koordinierte Reaktion erfordert. Diese Definition impliziert stabile Systemgrenzen, reproduzierbares Verhalten und klar zuordenbare Ursachen. Genau diese Annahmen sind bei KI-Systemen jedoch nur eingeschränkt gültig.

KI-Systeme zeichnen sich durch probabilistisches Verhalten, starke Kontextabhängigkeit und eine enge Kopplung an Datenverteilungen aus. Abweichungen vom erwarteten Verhalten sind nicht notwendigerweise Indikatoren für Defekte, sondern oft inhärenter Bestandteil der Systemfunktion. Gleichzeitig können scheinbar harmlose Abweichungen Vorboten schwerwiegender Vorfälle sein, insbesondere wenn sie sich unter Skalierung oder in veränderten Nutzungskontexten verstärken. Eine Incident-Definition, die entweder jede Abweichung eskaliert oder erst auf manifeste Schäden reagiert, ist für KI-Systeme gleichermaßen ungeeignet.

Hinzu kommt, dass Ursachen und Verantwortlichkeiten bei KI-Systemen häufig über Organisations- und Systemgrenzen hinweg verteilt sind. Trainingsdaten, Modelle, Deployments, Nutzungskontexte und regulatorische Rahmenbedingungen greifen ineinander, ohne dass ein einzelner Akteur das System vollständig kontrolliert. Eine praxistaugliche Incident-Definition muss dieser Verteilung Rechnung tragen und

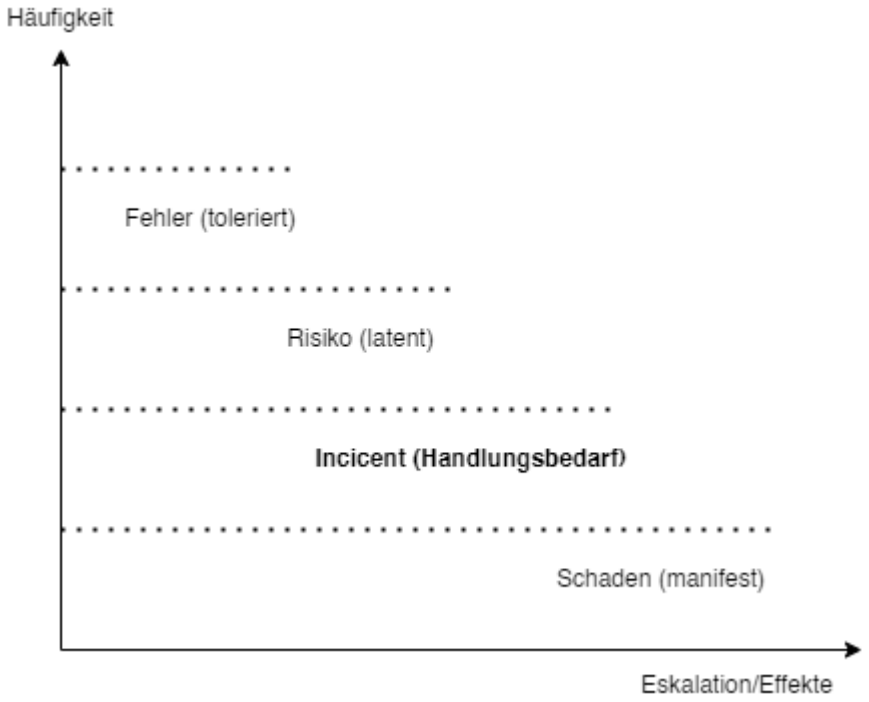
sich am beobachtbaren Verhalten sowie am daraus resultierenden Handlungsbedarf orientieren, nicht an idealisierten Annahmen über Systemkontrolle oder Fehlerfreiheit.

Eine eigenständige Incident-Definition für KI ist daher kein semantischer Luxus, sondern eine operative Notwendigkeit. Sie bildet die Grundlage für Entscheidungsfindung unter Zeitdruck, für Priorisierung begrenzter Ressourcen und für die Anschlussfähigkeit an Governance- und Meldepflichten. Ohne eine solche Definition bleibt Incident Response entweder reaktiv, übervorsichtig oder wirkungslos.

1.2 Fehler, Risiko und Incident – begriffliche Abgrenzungen

Die klare Unterscheidung zwischen Fehlern, Risiken und Incidents ist zentral für eine funktionierende KI Incident Response. Diese Begriffe werden in der Praxis häufig synonym verwendet, bezeichnen jedoch unterschiedliche Phänomene mit jeweils eigenen Konsequenzen für den Betrieb von KI-Systemen.

Ein Fehler beschreibt zunächst eine Abweichung vom gewünschten oder erwarteten Systemverhalten. Bei KI-Systemen kann dies ein falsches Klassifikationsergebnis, eine inkonsistente Antwort oder eine unplausible Empfehlung sein. Solche Fehler sind inhärent statistischen Modellen und stellen für sich genommen keinen Incident dar. Würden sie als solche behandelt, wäre ein stabiler Betrieb nicht möglich, da probabilistische Systeme notwendigerweise mit Fehlerraten arbeiten.



Einordnung von Fehlern, Risiken, Incidents und Schäden als aufeinanderfolgende Systemzustände. Incident Response setzt dort an, wo Kontrollannahmen über das Systemverhalten verletzt werden, unabhängig vom Eintritt eines Schadens.

Ein Risiko hingegen ist zukunftsgerichtet. Es beschreibt die Möglichkeit, dass ein bestimmtes Systemverhalten unter bestimmten Bedingungen zu Schaden führen kann. Risiken existieren unabhängig davon, ob sie sich jemals realisieren, und sind Gegenstand kontinuierlicher Bewertung und Steuerung. Incident Response setzt nicht bei Risiken an sich

an, sondern bei deren konkreter Manifestation oder unmittelbarer Eskalationsgefahr.

Ein Incident liegt vor, wenn sich ein Risiko in einem konkreten Ereignis oder einer Ereignisfolge materialisiert, die eine aktive Reaktion erfordert. Entscheidend ist dabei nicht zwingend der Eintritt eines Schadens, sondern der Verlust von Kontrolle über das Systemverhalten oder die Verletzung zentraler Annahmen, auf denen Betrieb, Sicherheit oder Compliance beruhen. Ein Incident kann somit auch dann vorliegen, wenn noch kein Schaden entstanden ist, die Fortsetzung des beobachteten Verhaltens jedoch nicht verantwortbar wäre.

Diese begriffliche Trennung erlaubt es, Incident Response gezielt dort einzusetzen, wo sie ihren größten Nutzen entfaltet: zwischen alltäglichen Abweichungen, die toleriert werden können, und manifesten Schäden, die bereits eingetreten sind. Sie bildet damit die Grundlage für eine abgestufte, verhältnismäßige und lernfähige Reaktion auf Vorfälle in KI-Systemen.

Für den weiteren Verlauf des Buches gilt folgende Arbeitsdefinition:

Ein KI-Incident ist ein beobachtbares Ereignis oder eine Serie von Ereignissen, bei denen das Verhalten eines KI-Systems außerhalb der vorgesehenen, akzeptierten oder regulatorisch zulässigen Grenzen liegt und eine Reaktion zur Begrenzung, Analyse oder Korrektur erforderlich macht.

Aus der begrifflichen Abgrenzung von Fehlern, Risiken und Incidents ergibt sich zwangsläufig ein Perspektivwechsel: Der KI-Incident ist

Stichwortverzeichnis

F

G

H

I

K

L

M

N

O

P

R

S

T

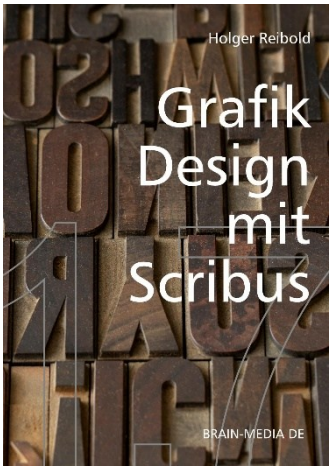
U

V

W

Z

Mehr von Brain-Media.de



Grafikdesign mit Scribus

In diesem Handbuch erfahren Sie alles, um mit Scribus ein professionelles Projekt umzusetzen – angefangen bei der Entwicklung kreativer Ideen bis zur konkreten Gestaltung.

Preis: 24,99 EUR

Umfang: 420 Seiten

Virtuelle Maschinen mit VirtualBox 7.x

So verwandeln Sie einen Rechner in ein ganzes Netzwerk oder bauen ein Testumgebung auf. Dieses Handbuch führt Sie in alle wichtigen Funktionen bis hin zur Cloud-Nutzung ein.

Preis: 16,99 EUR

Umfang: 150 Seiten

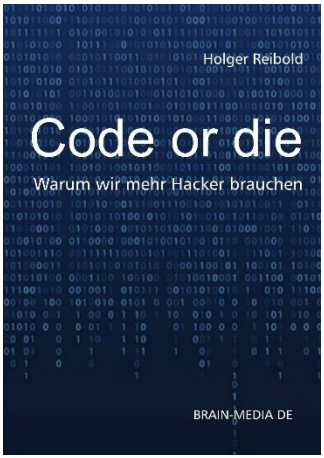
Audio Editing mit Audacity 4.x

Alles Wichtige, was Sie für den erfolgreichen Einsatz des freien Audioeditors wissen müssen.

Umfang: 220 Seiten

Preis: 19,99 EUR

Erscheint: Frühjahr 2026



Code or die

Ein Manifest für mehr digitale Selbstbestimmung, Neugierde und Eigenverantwortung. Medienkompetenzen alleine genügen nicht; die Gesellschaft von morgen braucht Digitalkompetenzen.

Umfang: 120 Seiten

Preis: 14,99 EUR

Erscheint Frühjahr 2026

Private KI – KI-Systeme lokal betreiben, kontrollieren und verantworten

Alles Wichtige für den sicheren Einsatz von lokalen KI-Systemen.

Umfang: 140 Seiten

Preis: 16,99 EUR

Erscheint: Februar 2026

KI-Sicherheit

Sichere KI ist eine Illusion – kontrollierbare KI ist ein Handwerk. Dieses Buch lehrt dieses Handwerk für die Praxis, jenseits von theoretischen Risikomodellen.

Umfang: 130 Seiten

Preis: 16,99 EUR

Erschienen: 03.01.2026