



Holger Reibold

KI Incident Response

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

BRAIN-MEDIA.DE

Holger Reibold

KI Incident Response

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

BRAIN-MEDIA.DE

Alle Rechte vorbehalten. Ohne ausdrückliche, schriftliche Genehmigung des Verlags ist es nicht gestattet, das Buch oder Teile daraus in irgendeiner Form durch Fotokopien oder ein anderes Verfahren zu vervielfältigen oder zu verbreiten. Dasselbe gilt auch für das Recht der öffentlichen Wiedergabe. Der Verlag macht darauf aufmerksam, dass die genannten Firmen- und Markennamen sowie Produktbezeichnungen in der Regel marken-, patent- oder warenrechtlichem Schutz unterliegen.

Verlag und Autor übernehmen keine Gewähr für die Funktionsfähigkeit beschriebener Verfahren und Standards.

© 2026 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

Inhaltsverzeichnis	I
Prolog	1
Vorwort	5
1 Was ist ein KI-Incident?.....	7
1.1 Warum eine präzise Definition notwendig ist.....	8
1.2 Fehler, Risiko und Incident – begriffliche Abgrenzungen	9
1.3 Near Misses als Frühindikatoren	14
1.4 Typische Klassen von KI-Incidents.....	15
1.5 Warum KI-Incidents schwer zu erkennen sind.....	18
1.6 Abgrenzung zu ethischen und politischen Kontroversen	20
1.7 Zwischenfazit.....	22
2 KI-Systeme als sozio-technische Systeme.....	25
2.1 Der KI-Lifecycle als sozio-technischer Prozess.....	26
2.2 Verteilte Verantwortung im KI-Lifecycle	28
2.3 Sozio-technische Kopplungen und ihre Auswirkungen	30
2.4 Konsequenzen für KI Incident Response	32
2.5 KI-Systeme unter realen Bedingungen.....	34
3 Risiko-, Threat- und Incident-Modelle für KI-Systeme	37
3.1 Vom Risiko zum Incident.....	38

3.2	Threat Modeling für KI-Systeme	41
3.3	Technische Angriffsklassen auf KI-Systeme	43
3.4	Incidents ohne klassischen Angreifer	45
3.5	Von Threats zu Incident-Klassen	46
3.6	Grenzen von Taxonomien und Modellen	48
3.7	Modelle als Werkzeuge, nicht als Wahrheit	49
4	Klassische Incident Response als Fundament.....	51
4.1	Beobachtbarkeit und Signalquellen	52
4.2	Detection von KI-Incidents	53
4.3	Erste Analyse unter Unsicherheit.....	55
4.4	Priorisierung und Severity-Einschätzung	56
4.5	Eskalation, Entscheidungsfindung und Kommunikation	57
4.6	Dokumentation und Übergang zur Response	59
5	KI Incident Response: Prinzipien und Ziele.....	61
5.1	Incident Response im KI-Kontext	62
5.2	Abgrenzung zu klassischer IT- und Security-IR.....	63
5.3	Ziele von KI Incident Response	64
5.4	Verhältnismäßigkeit und Eingriffstiefe	65
5.5	Grenzen technischer Interventionen	66
6	Rollen, Verantwortlichkeiten und Entscheidungsstrukturen	69
6.1	Operative Rollen im Incident-Fall	70

6.2	Frage der Verantwortung	71
6.3	Entscheidungsfindung unter Unsicherheit.....	72
6.4	Eskalationspfade und Stop-Kriterien	73
6.5	Umgang mit Verantwortungslücken.....	74
7	Technische Response-Maßnahmen	77
7.1	Sofortmaßnahmen im laufenden Betrieb.....	78
7.2	Input- und Output-Filterung	79
7.3	Zugriffsbeschränkungen und Rate Limiting	81
7.4	Temporäre Deaktivierung und Fallbacks	82
7.5	Risiken sekundärer Effekte	84
8	Deployment Corrections und Systemanpassungen	87
8.1	Korrekturen ohne Retraining	88
8.2	Prompt-, Kontext- und Policy-Anpassungen.....	89
8.3	Änderungen an Systemarchitektur und Tooling.....	91
8.4	Validierung von Korrekturmaßnahmen.....	92
8.5	Wann Korrekturen neue Incidents erzeugen	94
9	Retraining, Fine-Tuning und Modellwechsel.....	97
9.1	Wann Retraining sinnvoll ist – und wann nicht	98
9.2	Datenänderungen als Intervention.....	99
9.3	Risiken von Overfitting und Regression	100
9.4	Modellwechsel als Response-Strategie	102

9.5	Nachweis der Wirksamkeit.....	103
10	Kommunikation während KI-Incidents.....	105
10.1	Interne Kommunikation und Lagebilder	106
10.2	Kommunikation mit Management und Governance.....	107
10.3	Externe Kommunikation und Stakeholder	108
10.4	Transparenz versus Risiko	110
10.5	Kommunikation als Incident-Faktor.....	111
11	Dokumentation, Logging und Nachvollziehbarkeit.....	113
11.1	Anforderungen an Incident-Dokumentation.....	114
11.2	Technische und organisatorische Logfiles.....	115
11.3	Rekonstruktion von Entscheidungsprozessen	116
11.4	Grenzen der Nachvollziehbarkeit bei KI	117
11.5	Dokumentation als Governance-Instrument.....	118
12	Post-Incident-Analyse und Lernen.....	121
12.1	Vom Vorfall zur strukturellen Erkenntnis	122
12.2	Blameless Postmortems für KI-Systeme.....	123
12.3	Wiederholungsmuster und systemische Schwächen	124
12.4	Rückkopplung in Design und Training	125
12.5	Lernen unter regulatorischen Rahmenbedingungen	126
13	Integration in Risikomanagement und Governance	129
13.1	Incident Response als Teil des KI-Risikomanagements....	130

13.2	Model Cards, Risk Assessments und Audits.....	131
13.3	Steuerung über Policies und Standards.....	132
13.4	Rollen von Boards und Gremien.....	134
13.5	Governance jenseits von Checklisten	135
14	Regulatorische Anforderungen.....	137
14.1	Überblick relevanter Regulierungsansätze	138
14.2	Incident Reporting und Fristen.....	139
14.3	Spannungsfeld Technik-Recht.....	140
14.4	Dokumentations- und Nachweispflichten	142
14.5	Incident Response als Compliance-Fähigkeit	143
15	KI-IR in unterschiedlichen Domänen	145
15.1	Hochrisiko-Anwendungen	146
15.2	Verbraucher- und Content-nahe Systeme.....	147
15.3	Interne Entscheidungsunterstützung	149
15.4	Plattformen und Basismodelle	150
15.5	Domänenspezifische Trade-offs.....	151
16	Organisatorische Reife und Capability Building	153
16.1	Reifegradmodelle für KI Incident Response.....	154
16.2	Aufbau von Teams und Kompetenzen	155
16.3	Übungen und Simulationen	157
16.4	Metriken für Wirksamkeit.....	158

16.5	Von Ad-hoc-Reaktion zu etablierter Praxis.....	159
17	Grenzen, Kosten und Nebenwirkungen	161
17.1	Überreaktion und Systemverzerrung	162
17.2	False Positives und Vertrauensverlust	163
17.3	Ökonomische und organisatorische Kosten	164
17.4	Wann Nicht-Eingreifen rational ist.....	166
17.5	Incident Response als Balanceakt.....	167
18	Zukünftige Entwicklungen.....	169
18.1	Zunehmende Autonomie von KI-Systemen.....	170
18.2	Agentische Systeme und neue Incident-Typen	171
18.3	Automatisierte Incident Detection und Response	173
18.4	Grenzen der Automatisierung	174
18.5	Offene Forschungs- und Praxisfragen.....	175
19	Schlussbetrachtungen	177
19.1	Rückblick auf zentrale Konzepte.....	178
19.2	Incident Response als Normalfall	179
19.3	Verantwortung unter Unsicherheit	180
19.4	Von Vorfällen zu Vertrauen.....	182
19.5	Ausblick	183
	Epilog	185

Anhang A – Begriffsdefinitionen	IX
Anhang B – Taxonomie von KI-Incidents	XV
Anhang C – Framework-Mapping.....	XIX
Anhang D – Referenzprozess für KI Incident Response.....	XXI
Literatur- und Quellenverzeichnis	XXV
Stichwortverzeichnis	XXVII
Mehr von Brain-Media.de	XXXIII

Prolog

KI-Sicherheitsvorfälle sind kein Randphänomen. Sie sind kein Zukunfts-thema, kein Forschungsproblem und kein „Edge Case“ für besonders innovative Organisationen. Sie sind Realität – heute, in produktiven Systemen, mit realen Auswirkungen auf Menschen, Märkte und Ver-trauen. Vielfach wird KI Incident Response primär als operative Disziplin bewertet: Erkennen, Analysieren, Eindämmen und Beheben von Vorfäl-len in KI-Systemen. Doch neuere Literatur machen deutlich, dass dieser Blick zu kurz greift. KI Incident Response entwickelt sich zu einem integ-ralen 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-Vor-fälle 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 angeschlossfähige 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 Vorfälle 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

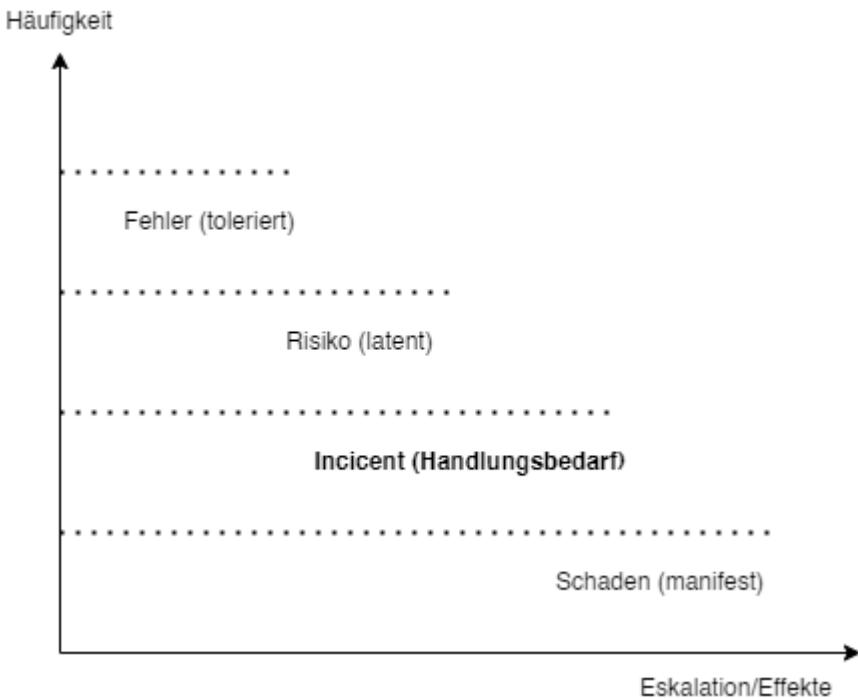
sich am beobachtbaren Verhalten sowie am daraus resultierenden Handlungsbedarf orientieren, nicht an idealisierten Annahmen über Systemkontrolle oder Fehlerfreiheit.

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

A

Adaptivität	42
Ad-hoc-Reaktion	159
Adversarialer Input	47
Agentische Systeme	171
Analyse	55
Angreifer	45
Angriffsfläche	41
Angriffsklasse	43, 44
Audit	131
Aufsicht	137
Aufsichtsgremium	134
Automatisierung	173
Autonomie	170

B

Balanceakt	167
Beobachtbarkeit	52
Beobachtung	114
Betrieb	26, 27
Blameless Postmortems	123
Board	134

C

Capability Building	153
Checkliste	135
Compliance	17, 143

D

Datenänderung	99
Deaktivierung	82
Defekt	7
Definition	8
Deployment	2, 5, 26, 27
Deployment Correction	87
Design	26
Designphase	26
Detection	48, 53
Determinismus	18
Dokumentation	59, 113
Dokumentationsanforderung	137
Dokumentationspflicht	142
Domäne	145
Domänenpezifisches	151

E

Eingriffstiefe	65
----------------------	----

Einschränkungen	62
Entscheidungsfindung	57
Entscheidungsorgan	134
Entscheidungsprozess	116
Entscheidungsstruktur	69
Entscheidungsunterstützung	149
Eskalation	57
Eskalationsgefahr	11
Eskalationslogik	170
Eskalationspfad	73
Ethik	20
EU AI Act	4, XX
Externe Kommunikation	108

F

Fallback	83
False Positives	163
Fehler	9
Fehlerrate	9
Fehlverhalten	17
Filtermechanismus	79
Filterung	79
Fine-Tuning	97
Framework-Mapping	XIX

G

Governance	1, 17, 47, 107, 129
GPAI	4
Grenze	66

H

Handlungsfähigkeit	113
Hochrisiko	146

I

Implikation	118
Incident Response	1
Incident-Datenbank	1
Incident-Dokumentation	114
Incident-Klasse	46
Incident-Kommunikation	4
Indikator	39
Input	79
Integrität	8
Interne Kommunikation	106
Interpretation	114
Intervention	66
IT-System	51

K

Kausalität	141
KI-Risikomanagement	130
KI-System	5
Klassen	15
Klassifikation	9
Kommunikation	57, 105
Kompetenzaufbau	155
Kontext	89

Kontextabhängigkeit	18
Kontextmechanismus	62
Kontextualisierung	55
Kontrolle	151
Korrekturmaßnahmen	92
Kosten	161
Künstliche Intelligenz	5

L

Lernen	126
Lernorientierung	126
Lifecycle	26
Logfiles	115
Logging	52

M

Manifestation	11
Meldepflicht	137
Metrik	52, 158
Model Card	131
Modell	25
Modellwechsel	97

N

Nachvollziehbarkeit	117
Nachweisanforderung	137
Nachweispflicht	142
Near Misses	1, 14

Nebenwirkung	161
Nicht-Eingreifen	166
NIST AI Risk Management Framework2	
NIST SP 800-61r3	2
Nutzungskontext	29

O

OECD	4
Offenheit	151
Operative Rolle	70
Organisatorische Reife	153
Output	30, 79
Overfitting	100
Ownership	69

P

Policy	89, 132
Post-Incident-Analyse	121
Priorisierung	56
Prompt	16, 89

R

Rate Limiting	81
Reale Bedingung	34
Rechtliches	126
Referenzprozess	XXI
Regression	100
Regulatorisches	126, 137

Regulierung	21
Regulierungsansatz	138
Reifegradmodell	154
Rekonstruktion	116
Resilienz.....	153
Response	48, 59
Response-Strategie	102
Retirement	26, 27
Retraining.....	97
Retrieval.....	44
Risiko	9, 38
Risikoklassifizierung	138
Risikomanagement.....	2, 129
Risikomodell.....	130
Risk Assessment	131
Rolle.....	69
Root Cause.....	56
Root-Cause-Analyse.....	33
Rückkopplung.....	125
Rückkopplungseffekt	30

S

Schadensbegrenzung	64
Selbstkorrektur.....	135
Severity	56
Signalquelle.....	52
Simulation	157
Skalierung	14, 81
Sofortmaßnahmen.....	78

Sozio-technisches System	25
Stakeholder.....	104, 109
Standard	132
Stop-Kriterien.....	73
Systemanpassung.....	87
Systemarchitektur.....	91
Systemgrenze	6
Systemkontext	47
Systemverzerrung	162

T

Taxonomie	37, XV
Technische Response	77
Telemetrie.....	20
These	67
Threat	37
Threat Modeling	41
Tooling.....	91
Training.....	26
Trainingsphase.....	26
Transparenz	110
Trigger.....	170

U

Überpriorisierung	57
Überreaktion	162
Unsicherheit	39, 72

V

Validierung	92
Verantwortlichkeit	69
Verantwortung	28, 71, 181
Verantwortungslücke	74
Verantwortungszuordnung	149
Verfügbarkeit	8
Verhältnismäßigkeit	65
Verteilung	19
Vertrauen	182
Vertrauensverlust	163
Vertraulichkeit	8

Verzerrung	19
Vorfall	122

W

Wertschöpfungskette	1
Wiederholungsmuster	124
Wirksamkeit	64, 103, 158
Wirksamkeitsnachweis	104

Z

Ziel	64
Zugriffsbeschränkung	81

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

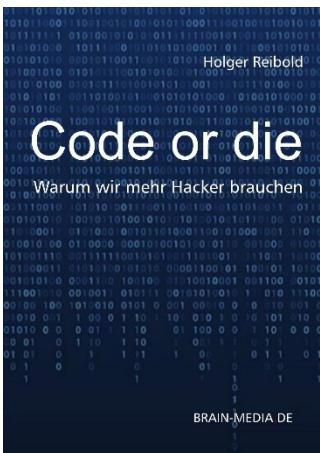
Audacity 4.x

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