

Holger Reibold

KI Incident Responce

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

Brain-Media.de

Dr. Holger Reibold – Hubert-Müller-Str. 52c – 66113 Saarbrücken

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

Inhaltsverzeichnis

Inhaltsverzeichnis	I
Vorwort	1
1 Was ist ein KI-Incident?.....	5
1.1 Warum eine präzise Definition notwendig ist.....	6
1.2 Fehler, Risiko und Incident – begriffliche Abgrenzungen	7
1.3 Near Misses als Frühindikatoren	12
1.4 Typische Klassen von KI-Incidents.....	13
1.5 Warum KI-Incidents schwer zu erkennen sind	16
1.6 Abgrenzung zu ethischen und politischen Kontroversen	18
1.7 Zwischenfazit.....	20
2 KI-Systeme als sozio-technische Systeme.....	23
2.1 Der KI-Lifecycle als sozio-technischer Prozess.....	24
2.2 Verteilte Verantwortung im KI-Lifecycle	26
2.3 Sozio-technische Kopplungen und ihre Auswirkungen	28
2.4 Konsequenzen für KI Incident Response	30
2.5 KI-Systeme unter realen Bedingungen	32
3 Risiko-, Threat- und Incident-Modelle für KI-Systeme.....	35
3.1 Vom Risiko zum Incident.....	36
3.2 Threat Modeling für KI-Systeme	39

3.3	Technische Angriffsklassen auf KI-Systeme	41
3.4	Incidents ohne klassischen Angreifer	43
3.5	Von Threats zu Incident-Klassen	44
3.6	Grenzen von Taxonomien und Modellen	46
3.7	Modelle als Werkzeuge, nicht als Wahrheit	47
4	Klassische Incident Response als Fundament.....	49
4.1	Beobachtbarkeit und Signalquellen	50
4.2	Detection von KI-Incidents	51
4.3	Erste Analyse unter Unsicherheit.....	53
4.4	Priorisierung und Severity-Einschätzung	54
4.5	Eskalation, Entscheidungsfindung und Kommunikation	55
4.6	Dokumentation und Übergang zur Response	57
5	KI Incident Response: Prinzipien und Ziele.....	59
5.1	Incident Response im KI-Kontext	60
5.2	Abgrenzung zu klassischer IT- und Security-IR.....	61
5.3	Ziele von KI Incident Response	62
5.4	Verhältnismäßigkeit und Eingriffstiefe	63
5.5	Grenzen technischer Interventionen	64
6	Rollen, Verantwortlichkeiten und Entscheidungsstrukturen.....	67
6.1	Operative Rollen im Incident-Fall	68
6.2	Frage der Verantwortung	69

6.3	Entscheidungsfindung unter Unsicherheit.....	70
6.4	Eskalationspfade und Stop-Kriterien	71
6.5	Umgang mit Verantwortungslücken.....	72
7	Technische Response-Maßnahmen	75
7.1	Sofortmaßnahmen im laufenden Betrieb.....	76
7.2	Input- und Output-Filterung	77
7.3	Zugriffsbeschränkungen und Rate Limiting	79
7.4	Temporäre Deaktivierung und Fallbacks	80
7.5	Risiken sekundärer Effekte	82
8	Deployment Corrections und Systemanpassungen	85
8.1	Korrekturen ohne Retraining	86
8.2	Prompt-, Kontext- und Policy-Anpassungen	87
8.3	Änderungen an Systemarchitektur und Tooling.....	89
8.4	Validierung von Korrekturmaßnahmen.....	90
8.5	Wann Korrekturen neue Incidents erzeugen	92
9	Retraining, Fine-Tuning und Modellwechsel.....	95
9.1	Wann Retraining sinnvoll ist – und wann nicht	96
9.2	Datenänderungen als Intervention.....	97
9.3	Risiken von Overfitting und Regression	98
9.4	Modellwechsel als Response-Strategie	100
9.5	Nachweis der Wirksamkeit.....	101

10	Kommunikation während KI-Incidents.....	103
10.1	Interne Kommunikation und Lagebilder	104
10.2	Kommunikation mit Management und Governance.....	105
10.3	Externe Kommunikation und Stakeholder	106
10.4	Transparenz versus Risiko	108
10.5	Kommunikation als Incident-Faktor	109
11	Dokumentation, Logging und Nachvollziehbarkeit.....	111
11.1	Anforderungen an Incident-Dokumentation	112
11.2	Technische und organisatorische Logfiles.....	113
11.3	Rekonstruktion von Entscheidungsprozessen	114
11.4	Grenzen der Nachvollziehbarkeit bei KI	115
11.5	Dokumentation als Governance-Instrument.....	116
12	Post-Incident-Analyse und Lernen.....	119
12.1	Vom Vorfall zur strukturellen Erkenntnis	120
12.2	Blameless Postmortems für KI-Systeme.....	121
12.3	Wiederholungsmuster und systemische Schwächen	122
12.4	Rückkopplung in Design und Training	123
12.5	Lernen unter regulatorischen Rahmenbedingungen	124
13	Integration in Risikomanagement und Governance	127
13.1	Incident Response als Teil des KI-Risikomanagements.....	128
13.2	Model Cards, Risk Assessments und Audits.....	129

13.3	Steuerung über Policies und Standards.....	130
13.4	Rollen von Boards und Gremien.....	132
13.5	Governance jenseits von Checklisten.....	133
14	Regulatorische Anforderungen.....	135
14.1	Überblick relevanter Regulierungsansätze	136
14.2	Incident Reporting und Fristen.....	137
14.3	Spannungsfeld Technik-Recht.....	138
14.4	Dokumentations- und Nachweispflichten	140
14.5	Incident Response als Compliance-Fähigkeit	141
15	KI-IR in unterschiedlichen Domänen	143
15.1	Hochrisiko-Anwendungen	144
15.2	Verbraucher- und Content-nahe Systeme.....	145
15.3	Interne Entscheidungsunterstützung.....	147
15.4	Plattformen und Basismodelle	148
15.5	Domänenspezifische Trade-offs.....	149
16	Organisatorische Reife und Capability Building	151
16.1	Reifegradmodelle für KI Incident Response.....	152
16.2	Aufbau von Teams und Kompetenzen	153
16.3	Übungen und Simulationen	155
16.4	Metriken für Wirksamkeit.....	156
16.5	Von Ad-hoc-Reaktion zu etablierter Praxis.....	157

17	Grenzen, Kosten und Nebenwirkungen	159
17.1	Überreaktion und Systemverzerrung	160
17.2	False Positives und Vertrauensverlust	161
17.3	Ökonomische und organisatorische Kosten	162
17.4	Wann Nicht-Eingreifen rational ist.....	164
17.5	Incident Response als Balanceakt	165
18	Zukünftige Entwicklungen.....	167
18.1	Zunehmende Autonomie von KI-Systemen.....	168
18.2	Agentische Systeme und neue Incident-Typen	169
18.3	Automatisierte Incident Detection und Response	171
18.4	Grenzen der Automatisierung	172
18.5	Offene Forschungs- und Praxisfragen	173
19	Schlussbetrachtungen	175
19.1	Rückblick auf zentrale Konzepte.....	176
19.2	Incident Response als Normalfall	177
19.3	Verantwortung unter Unsicherheit	178
19.4	Von Vorfällen zu Vertrauen.....	180
19.5	Ausblick	181
	Zum Schluss	183
	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

Vorwort

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

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 Kontrollmaßnahmen 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

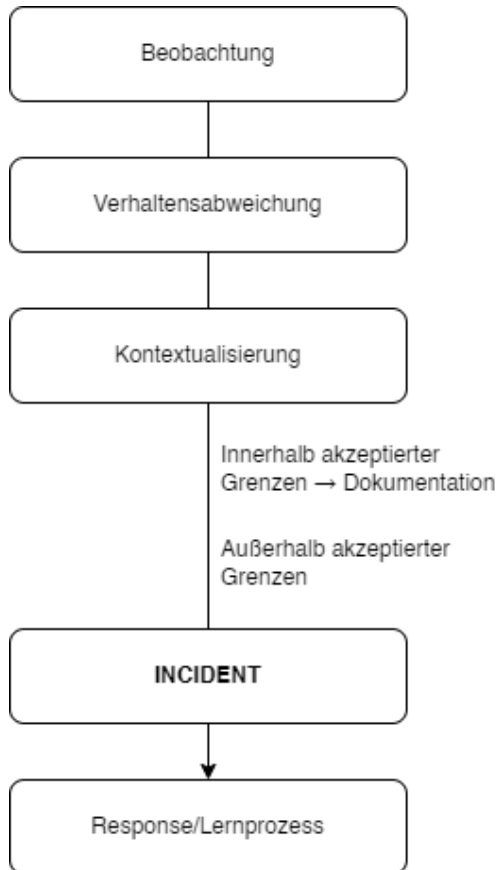
weniger als technischer Defekt zu verstehen, sondern primär als verhaltensbasierter Handlungsanlass. Im Zentrum steht nicht die Frage, ob ein Modell korrekt implementiert wurde oder ob seine statistischen Eigenschaften formal den Erwartungen entsprechen, sondern ob das beobachtbare Systemverhalten unter realen Einsatzbedingungen weiterhin kontrolliert und verantwortbar ist.

Diese Sichtweise trägt der Tatsache Rechnung, dass KI-Systeme häufig auch dann „funktionieren“, wenn sie Incidents verursachen. Ein Sprachmodell kann formal korrekte, grammatikalisch einwandfreie Antworten liefern und dennoch sicherheitsrelevante oder regulatorisch problematische Inhalte erzeugen. Ein Empfehlungssystem kann innerhalb seiner spezifizierten Metriken optimieren und zugleich systematisch unerwünschte Effekte verstärken. In solchen Fällen liegt kein klassischer Defekt vor, wohl aber ein Incident im Sinne eines Kontrollverlusts über das Systemverhalten.

Ein KI-Incident ist daher immer kontextabhängig. Die Grenze dessen, was als akzeptables Verhalten gilt, ergibt sich aus dem vorgesehenen Einsatzzweck, den getroffenen Annahmen über Nutzung und Missbrauch, sowie aus rechtlichen und organisatorischen Rahmenbedingungen. Dieselbe Systemreaktion kann in einem Kontext tolerierbar und in einem anderen nicht akzeptabel sein. Incident Response muss diese Kontextabhängigkeit explizit berücksichtigen, anstatt nach universellen Schwellenwerten zu suchen.

Die Konsequenz: Die Incident Detection kann nicht ausschließlich auf Schadensindikatoren oder bekannte Angriffsmuster ausgerichtet sein.

Sie muss in der Lage sein, Abweichungen vom erwarteten Verhaltensraum zu erkennen, auch wenn diese noch keine unmittelbaren negativen Effekte zeigen. Der KI-Incident markiert damit den Punkt, an dem Beobachtung in Handlung übergeht – nicht notwendigerweise den Punkt maximaler Eskalation.



Der Übergang von der Beobachtung einer Verhaltensabweichung zur Incident Response. Der Incident markiert den Punkt, an dem Analyse in operative Intervention übergeht.

Literatur- und Quellenverzeichnis

Center for Security and Emerging Technology (2022). AI incidents: An emerging risk (CSET Issue Brief). Georgetown University.

European Commission (2024). Artificial Intelligence Act: Implementation, technical neutrality and governance (AI Act Implementation Forum Report).

European Union (2024). Regulation (EU) 2024/... laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.

National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). NIST.

National Institute of Standards and Technology (2024). Computer Security Incident Handling Guide (Special Publication 800-61 Revision 3). NIST.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crutchfield, J. P., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E.,

- Shariff, A., ... Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
- Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., & Amodei, D. (2023). Deployment corrections: An incident response framework for frontier AI models. *arXiv preprint arXiv:2302.04844*.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504.
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752.
- World Health Organization (2021). *Ethics and governance of artificial intelligence for health*. WHO Press.
- Zhou, Q., Danks, D., & London, A. J. (2022). Towards a taxonomy of AI risks in the health domain. *Journal of Biomedical Informatics*, 126, 103996.

Stichwortverzeichnis

A

Adaptivität	40
Ad-hoc-Reaktion.....	157
Adversarialer Input	45
Agentische Systeme	169
Analyse	53
Angreifer.....	43
Angriffsfläche	39
Angriffsklasse.....	41, 42
Audit	129
Aufsicht.....	135
Aufsichtsgremium	132
Automatisierung	171
Autonomie.....	168

B

Balanceakt	165
Beobachtbarkeit	50
Beobachtung	112
Betrieb.....	24, 25
Blameless Postmortems.....	121
Board	132

C

Capability Building	151
Checkliste	133
Compliance	15, 141

D

Datenänderung	97
Deaktivierung.....	80
Defekt	5
Definition.....	6
Deployment.....	2, 24, 25
Deployment Correction.....	85
Design	24
Designphase	24
Detection	46, 51
Determinismus	16
Dokumentation	57, 111
Dokumentationsanforderung ...	135
Dokumentationspflicht.....	140
Domäne.....	143
Domänenspezifisches	149

E

Eingriffstiefe	63
Einschränkungen	60
Entscheidungsfindung.....	55
Entscheidungsorgan.....	132
Entscheidungsprozess.....	114
Entscheidungsstruktur.....	67
Entscheidungsunterstützung ...	147
Eskalation	55
Eskalationsgefahr.....	9
Eskalationslogik	168
Eskalationspfad	71
Ethik	18
EU AI Act.....	4, XX
Externe Kommunikation.....	106

F

Fallback.....	81
False Positives.....	161
Fehler	7
Fehlerrate.....	7
Fehlverhalten.....	15
Filtermechanismus	77
Filterung.....	77
Fine-Tuning	95
Framework-Mapping	XIX

G

Governance.....	1, 15, 45, 105, 127
GPAI.....	4
Grenze	64

H

Handlungsfähigkeit	111
Hochrisiko	144

I

Implikation	116
Incident Response	1
Incident-Datenbank	1
Incident-Dokumentation	112
Incident-Klasse.....	44
Incident-Kommunikation	4
Indikator	37
Input	77
Integrität.....	6
Interne Kommunikation	104
Interpretation	112
Intervention.....	64
IT-System	49

K

Kausalität	139
KI-Risikomanagement.....	128

Klassen.....	13
Klassifikation	7
Kommunikation.....	55, 103
Kompetenzaufbau	153
Kontext.....	87
Kontextabhängigkeit.....	16
Kontextmechanismus	60
Kontextualisierung.....	53
Kontrolle	149
Korrekturmaßnahmen	90
Kosten	159

L

Lernen.....	124
Lernorientierung	124
Lifecycle	24
Logfiles	113
Logging.....	50

M

Manifestation	9
Meldepflicht.....	135
Metrik.....	50, 156
Model Card	129
Modell.....	23
Modellwechsel	95

N

Nachvollziehbarkeit	115
Nachweisanforderung.....	135
Nachweispflicht	140
Near Misses	1, 12
Nebenwirkung	159
Nicht-Eingreifen.....	164
NIST AI Risk Management Framework	2
NIST SP 800-61	XX
NIST SP 800-61r3.....	2
Nutzungskontext.....	27

O

OECD	4
Offenheit	149
Operative Rolle	68
Organisatorische Reife	151
Output	28, 77
Overfitting	98
Ownership.....	67

P

Policy	87, 130
Post-Incident-Analyse	119
Priorisierung	54
Prompt.....	14, 87

R

Rate Limiting	79
Reale Bedingung	32
Rechtliches	124
Referenzprozess	XXI
Regression	98
Regulatorisches.....	124, 135
Regulierung	19
Regulierungsansatz	136
Reifegradmodell.....	152
Rekonstruktion	114
Resilienz.....	151
Response	46, 57
Response-Strategie.....	100
Retirement.....	24, 25
Retraining	95
Retrieval.....	42
Risiko	7, 36
Risikoklassifizierung	136
Risikomanagement	2, 127
Risikomodell.....	128
Risk Assessment	129
Rolle	67
Root Cause	54
Root-Cause-Analyse	31
Rückkopplung.....	123
Rückkopplungseffekt.....	28

S

Schadensbegrenzung.....	62
Selbstkorrektur	133
Severity	54
Signalquelle	50
Simulation.....	155
Skalierung	12, 79
Sofortmaßnahmen	76
Sozio-technisches System	23
Stakeholder	102, 107
Standard.....	130
Stop-Kriterien.....	71
Systemanpassung	85
Systemarchitektur	89
Systemkontext.....	45
Systemverzerrung	160

T

Taxonomie	35, XV
Technische Response.....	75
Telemetrie	18
These	65
Threat	35
Threat Modeling	39
Tooling	89
Training	24
Trainingsphase	24
Transparenz	108

Trigger 168

U

Überpriorisierung 55

Überreaktion..... 160

Unsicherheit 37, 70

V

Validierung 90

Verantwortlichkeit..... 67

Verantwortung26, 69, 179

Verantwortungslücke..... 72

Verantwortungszuordnung 147

Verfügbarkeit..... 6

Verhältnismäßigkeit..... 63

Verteilung 17

Vertrauen 180

Vertrauensverlust 161

Vertraulichkeit..... 6

Verzerrung 17

Vorfall 120

W

Wertschöpfungskette 1

Wiederholungsmuster 122

Wirksamkeit 62, 101, 156

Wirksamkeitsnachweis..... 102

Z

Ziel 62

Zugriffsbeschränkung 79



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

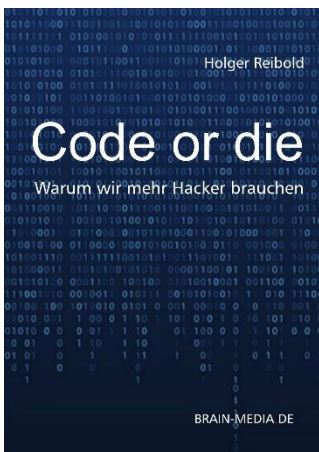
Audacity 4.x

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