Universität Augsburg
University

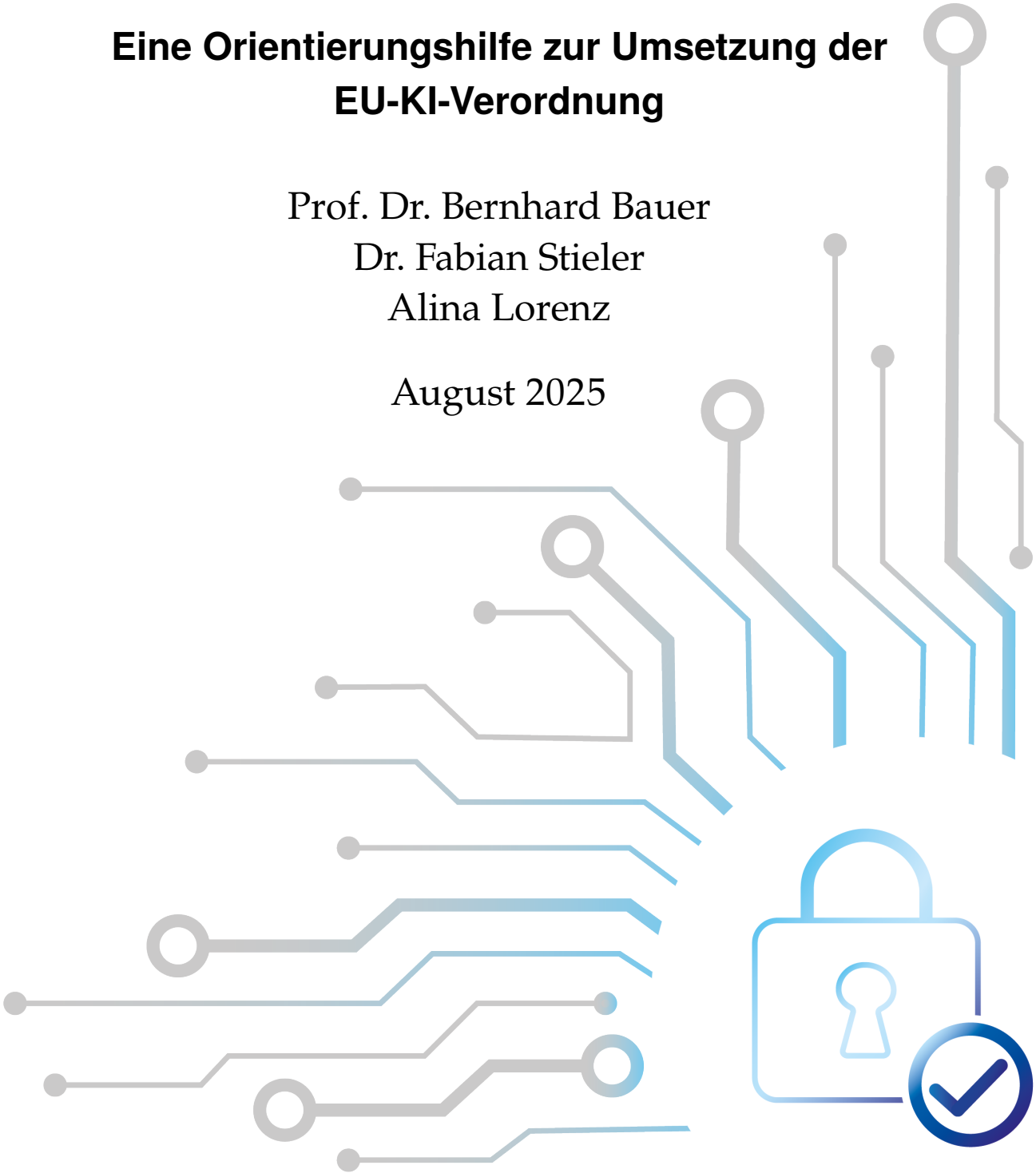
Audit- und Wissensplattform für Künstliche Intelligenz

Cybersicherheit in KI-Systemen

Eine Orientierungshilfe zur Umsetzung der EU-KI-Verordnung

Prof. Dr. Bernhard Bauer
Dr. Fabian Stieler
Alina Lorenz

August 2025



Executive Summary

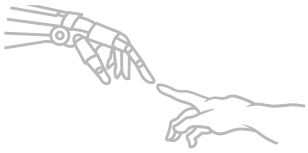
Die Cybersicherheit in KI-Systemen ist ein wesentlicher Aspekt zur Sicherstellung der Konformität von KI-Systemen gemäß den Anforderungen der EU-KI-Verordnung. Sie dient als Nachweis dafür, dass ein KI-System alle gesetzlichen Vorgaben erfüllt, sicher und robust läuft und begleitet dessen gesamten Lebenszyklus.

Wir geben in diesem Leitfaden einen Überblick über die Anforderungen an Cybersicherheit und zeigen, wie diese strukturiert und praktisch umgesetzt werden können, um die Entwicklung und den Betrieb verantwortungsvoller KI-Systeme zu gewährleisten.

Disclaimer
Teile dieses Textes wurden versuchsweise mit KI generiert und qualitätsgesichert, sowie die Lesbarkeit mithilfe von KI überarbeitet. Zur besseren Lesbarkeit wurde das generische Maskulinum verwendet. Die verwendeten Personenbezeichnungen beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Geschlechter.

Publisher
Audit- und Wissensplattform für vertrauenswürdige KI (AWIKI)
https://www.awiki.eu
August 2025

Inhaltsverzeichnis



1 Einleitung

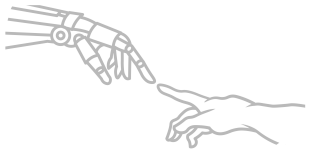
1.1 Hintergrund

Die europäische KI-Verordnung (KI-VO) fordert von KI-Systemen in Art. 15 (1) mit hohem Risiko einen komplexen regulatorischen Rahmen, der darauf abzielt, *Hochrisiko-KI-Systeme so zu konzipieren und zu entwickeln , dass sie ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit erreichen und in dieser Hinsicht während ihres gesamten Lebenszyklus beständig funktionieren* [1]. Wir konzentrieren uns in diesem Leitfaden auf Cybersicherheit, Betrachtungen bzgl. Robustheit sind nicht Teil dieses Leitfadens.

[1]: Das Europäische Parlament und der Rat der Europäischen Union (2024), *EU Artificial Intelligence Act*

Zu diesem Zweck verlangt die KI-VO in Art. 15 (5) im Hinblick auf **Cybersicherheit**, dass *Hochrisiko-KI-Systeme widerstandsfähig gegen Versuche unbefugter Dritter sein müssen, ihre Verwendung, Ausgaben oder Leistung durch Ausnutzung von Systemschwachstellen zu verändern. Die technischen Lösungen zur Gewährleistung der Cybersicherheit von Hochrisiko-KI-Systemen müssen den jeweiligen Umständen und Risiken angemessen sein. Die technischen Lösungen für den Umgang mit KI-spezifischen Schwachstellen umfassen gegebenenfalls Maßnahmen, um Angriffe, mit denen versucht wird, eine Manipulation des Trainingsdatensatzes („data poisoning“) oder vortrainierter Komponenten, die beim Training verwendet werden („model poisoning“), vorzunehmen, Eingabedaten, die das KI-Modell zu Fehlern verleiten sollen („adversarial examples“ oder „model evasions“), Angriffe auf vertrauliche Daten oder Modellmängel zu verhüten, zu erkennen, darauf zu reagieren, sie zu beseitigen und zu kontrollieren.*

Der vorliegende Leitfaden ist wie folgt strukturiert: Zunächst werden die Grundlagen der Cybersicherheit dargelegt und in Beziehung zu den rechtlichen Grundlagen der KI-VO, zu unterstützenden Normen sowie zu Sicherheitsrisiken entlang des Produktlebenszyklus gesetzt.



Im weiteren Verlauf werden die Sicherheitsanforderungen in den verschiedenen Produktlebenszyklusphasen dargelegt und Maßnahmen zur Verhinderung von Angriffen erörtert. Der Abschnitt schließt mit einer Zusammenfassung, wobei neben der klassischen Cybersicherheit, die auch für KI-Systeme relevant ist, der Schwerpunkt auf der spezifischen Cybersicherheit von KI-Systemen liegt. In diesem Zusammenhang werden unterschiedliche Lernverfahren und Modellarten diskutiert und hinsichtlich ihrer Sicherheit bewertet.

Abschließend wird aufgezeigt, wie die Ausführungen mit der Umsetzung der KI-VO korrelieren.

1.2 Motivation und Begriffsklärung

In den letzten Jahren hat die künstliche Intelligenz signifikante Fortschritte erzielt, wodurch die Zahl der Anwendungsbereiche stetig gewachsen ist. Beispiele hierfür sind die Medizin, autonome Fahrzeuge, die Finanzwelt und sogar der Einsatz von KI in der Cybersicherheit. Gleichzeitig ist jedoch auch eine Zunahme der Risiken festzustellen, dass diese Systeme zum Ziel von Angriffen oder Missbrauch werden.

Zur Begriffsklärung wollen wir zunächst verschiedene, aber verwandte Begrifflichkeiten klären, welche in Abbildung 1.1 aufgelistet sind.

Sicherheit (**Security**)

Security stellt einen übergreifenden Begriff dar, der alle Maßnahmen umfasst, die den fortlaufenden Betrieb und die kritischen Funktionen einer Organisation auch unter Bedrohungslagen sicherstellen. Diese Bedrohungen können sowohl technischer Natur sein – etwa durch Angriffe auf Informationssysteme – als auch physischer Herkunft, wie Naturkatastrophen oder Einbrüche. Sicherheitsmaßnahmen können dabei präventiver, detektiver oder korrektiver Natur sein und beinhalten Strategien wie Abschreckung, Vermeidung, Prävention, Erkennung, Wiederherstellung und Korrektur. Ziel ist

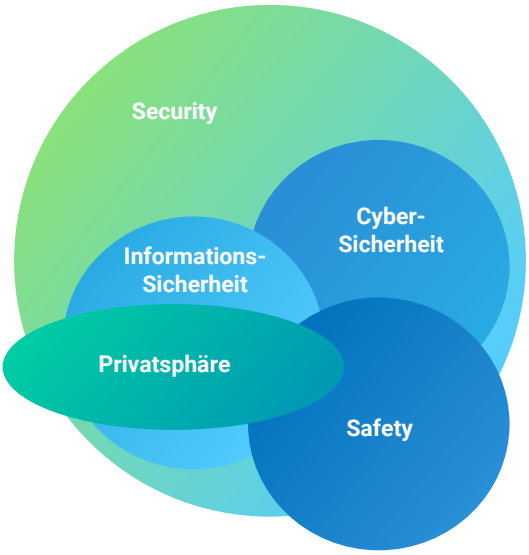


Abbildung 1.1: Begriffe und deren Zusammenhang

ein ganzheitlicher Ansatz, der sowohl technische als auch organisatorische Schutzmaßnahmen einschließt.

Cybersicherheit (Cybersecurity)

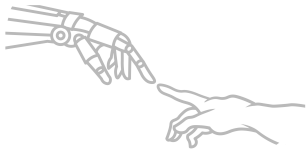
Im engeren Sinne versteht man unter Cybersecurity den Schutz von IT-Systemen und den darin enthaltenen Daten. Der Fokus liegt hierbei auf der Gewährleistung von Verfügbarkeit, Integrität, Authentizität, Vertraulichkeit und Nichtabstreitbarkeit digitaler Informationen. Cybersecurity ist somit auf digitale Systeme beschränkt und bezieht sich nicht auf analoge Informationen, wie z.B. physisch gespeicherte Verträge oder Personalakten.

Informationssicherheit

Informationssicherheit hingegen verfolgt einen umfassenderen Ansatz: Sie bezieht sich auf den Schutz von Informationen unabhängig vom Trägermedium. Dies bedeutet, dass sowohl digitale als auch analoge Informationsquellen – etwa Papierdokumente – unter den Schutzbereich fallen.

Privatsphäre (Privacy)

Ein eng verwandtes, aber dennoch eigenständiges Konzept ist die Privatsphäre. Diese bezieht sich auf den Schutz personenbezogener Informationen und gewährleistet, dass der Zugriff auf sensible Daten nur autorisierten Personen erlaubt ist. Ein typisches Beispiel hierfür ist der beschränkte Zugang zu Patientendaten in medizinischen Einrichtungen, welcher ausschließlich behandelnden Fachpersonen gestattet ist.



Safety

Abschließend ist Sicherheit im Sinne von Safety ein Begriff, der den Schutz vor physischen Schäden und Gefahren beschreibt. Dies umfasst Maßnahmen zur Verhinderung von Todesfällen, Verletzungen, beruflich bedingten Erkrankungen, Sachschäden sowie Umweltschäden. In Abgrenzung zur Cybersecurity liegt hier der Fokus auf dem physischen Wohl und der Unversehrtheit von Personen, Infrastruktur und Umwelt.

Ein konkretes Beispiel für das Zusammenwirken dieser Begriffe, das in Tabelle 1.1 dargestellt ist, ist der Ausfall oder die Schwäche kryptografischer Verfahren. Wird etwa ein Kommunikationskanal unzureichend verschlüsselt, entsteht dadurch zunächst ein Cybersecurity-Problem, da die Integrität und Vertraulichkeit der digital übermittelten Daten gefährdet ist. Die Angriffsfläche betrifft hier typischerweise digitale Infrastrukturen wie Netzwerke, Server oder Anwendungen.

Diese technische Schwachstelle führt jedoch über die rein informationstechnische Dimension hinaus: Wird durch den Mangel an Verschlüsselung der Zugriff auf sensible Daten ermöglicht – beispielsweise Gesundheitsdaten, Personaldaten oder Kundendaten –, liegt zugleich eine Verletzung der Informationssicherheit vor. Informationssicherheit bezieht sich nämlich nicht nur auf Systeme, sondern auf den Schutz des Inhalts selbst, unabhängig davon, ob dieser in digitaler oder analoger Form vorliegt.

Tabelle 1.1: Zusammenspiel der Konzepte - Beispielhafte Kaskade eines Vorfalls

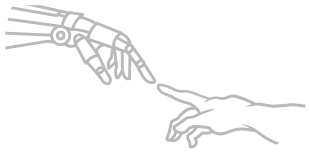
Ereignis	**Betroffener Bereich**	**Beschreibung**
Schwache Verschlüsselung eines Kommunikationskanals	Cybersecurity	Gefährdung der Vertraulichkeit & Integrität digitaler Daten
Zugriff auf sensible Informationen (z.B. Gesundheitsdaten)	Informationssicherheit	Schutz des Inhalts wird verletzt – unabhängig vom Medium
Zugriff auf personenbezogene Daten	Privacy	Verletzung der Privatsphäre – rechtliche & ethische Konsequenzen
Manipulation medizinischer Daten	Safety	Potenziell physische Gefährdung betroffener Personen

In Fällen, in denen die kompromittierten Daten personenbezogene Informationen enthalten, ist zusätzlich der Bereich der Privacy betroffen. Die Verletzung der Privatsphäre durch unautorisierten Zugriff auf personenbezogene Daten stellt nicht nur einen ethischen und rechtlichen Verstoß dar (z.B. gegen die DSGVO), sondern beeinträchtigt auch das Vertrauen von Individuen und Gesellschaft in datenverarbeitende Systeme.

Tabelle 1.2: Terminologie

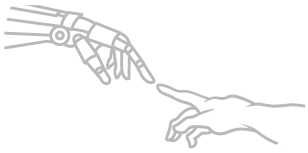
Begriff	Definition	Schutzobjekte	Fokus	Beispielhafte Bedrohungen	Typische Maßnahmen
Security	Übergreifender Begriff für Maßnahmen zum Schutz des Betriebs und der Funktionen einer Organisation unter Bedrohungslagen	Organisation Prozesse Infrastruktur	Ganzheitlich	Cyberangriffe Naturkatastrophen Einbruch	Abschreckung Prävention Erkennung Wiederherstellung
Cybersecurity	Schutz von IT-Systemen und digitalen Daten	Digitale Systeme & Informationen	Digital	Malware Hackerangriffe Datenlecks Adversarial Attacks Modellextraktion	Firewalls Verschlüsselung Zugriffskontrollen Anomalieerkennung
Informationssicherheit	Schutz von Informationen unabhängig vom Trägermedium	Analoge & digitale Informationen Trainingsdaten Inferenzdaten Modelle	Inhalt	Diebstahl von Dokumenten Systemkompromittierung Datenverfälschung Data Poisoning	Klassifizierung Zugangskontrolle Versionierung Backup
Privatsphäre (Privacy)	Schutz personenbezogener Daten vor unautorisiertem Zugriff	Personenbezogene Informationen	Individuum	Datenmissbrauch Unberechtigter Zugriff Deanonymisierung Unautorisierte Profilbildung	Datensparsamkeit Einwilligung Differential Privacy DSGVO-konforme Verarbeitung
Sicherheit (Safety)	Schutz vor physischen sowie psychischen Schäden und Gefahren	Menschen Umwelt Gesellschaft Sachwerte	Physisch	Unfälle Produktausfälle Umweltkatastrophen	Sicherheitsregeln Human-in-the-Loop Failsafe-Design Simulationsbasiertes Testing Notabschaltmechanismen

Unter bestimmten Umständen kann ein solcher Vorfall sogar Folgen im Bereich der Safety nach sich ziehen. Wird beispielsweise ein medizinisches Informationssys-



tem angegriffen und Patientendaten manipuliert oder offengelegt, kann dies physische Gesundheitsrisiken für betroffene Personen verursachen – etwa durch Fehlbehandlungen oder den Missbrauch sensibler Daten. In diesem Fall verschmelzen Fragen der digitalen Sicherheit mit solchen der physischen Unversehrtheit.

Die zuvor erläuterten Begriffe zeigen bereits, dass sich die unterschiedlichen Sicherheitsdimensionen nicht isoliert betrachten lassen, sondern in realen Szenarien häufig ineinandergreifen. Besonders im Kontext von KI-Systemen können einzelne Vorfälle – etwa ein technischer Fehler oder ein gezielter Angriff – mehrere dieser Schutzbereiche gleichzeitig betreffen. Um diese Wechselwirkungen greifbarer zu machen, wird Tabelle 1.2 bereitgestellt. Diese konkretisiert eine systematische Gegenüberstellung der Begrifflichkeiten und zählt sowohl klassische als auch KI-System spezifische Aspekte auf.



Grundlagen 2

In diesem Grundlagenkapitel werfen wir einen detaillierten Blick auf die rechtlichen Grundlagen, nämlich die Anforderungen aus der KI-VO, sowie unterstützende Standards für Cybersicherheit.

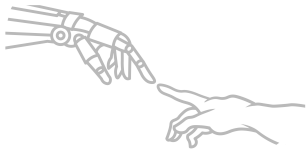
Darüber hinaus diskutieren wir Sicherheitsziele und skizzieren Angriffsarten spezifisch für KI-Systeme.

2.1 Rechtliche Grundlagen

Gemäß Artikel 15 der KI-VO [1] müssen KI-Systeme während ihres gesamten Lebenszyklus ein angemessenes Niveau an Genauigkeit, Robustheit und Cybersicherheit erreichen und dieses Niveau beibehalten. Die Genauigkeitsgrade und die entsprechenden Bewertungsparameter müssen in der Gebrauchsanweisung angegeben werden.

Sowohl Robustheit als auch Cybersecurity können durch technische Redundanzlösungen erreicht werden, die Backup- oder Fail-Safe-Pläne beinhalten können. Insbesondere müssen KI-Systeme so widerstandsfähig wie möglich gegenüber Fehlern, Störungen oder Unstimmigkeiten sein, die innerhalb des Systems oder in der Umgebung, in der das System betrieben wird, auftreten können, vor allem aufgrund ihrer Interaktion mit natürlichen Personen oder anderen Systemen. Das betrifft insbesondere auch KI-Systeme, die nach dem Inverkehrbringen oder der Inbetriebnahme weiterlernen, und das Risiko bergen, dass potenziell verzerrte Ergebnisse den Input für künftige Modelle beeinflussen.

Zu diesem Zweck sind Maßnahmen zur Verhinderung, Erkennung, Reaktion, Behebung und Kontrolle von Angriffen zu ergreifen, die beispielsweise darauf abzielen, den Trainingsdatensatz zu schützen. In diesen Kontext



fallen ebenfalls die Manipulation von vortrainierten KI-Modellen, die Ausführung von Eingaben, die das KI-Modell zu Fehlern veranlassen sollen, sowie Angriffe auf die Vertraulichkeit der einzelnen Komponenten oder des Gesamtsystems. Die genannten Angriffe können durch die Beeinflussung von Operationen, die Ausschließung oder Minimierung von Rückkopplungsschleifen sowie durch die Sicherstellung der ordnungsgemäßen Behandlung solcher Rückkopplungsschleifen durch geeignete Abhilfemaßnahmen verhindert oder minimiert werden.

Tabelle 2.1: Anforderungen der KI-VO im Bezug auf Robustheit und Cybersicherheit für KI-Systeme mit hohem Risiko.

Forderung	Quelle	Beschreibung
Genauigkeit Robustheit Cybersicherheit	Artikel 15 (1)	Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit erreichen und in dieser Hinsicht während ihres gesamten Lebenszyklus beständig funktionieren.
Widerstandsfähigkeit	Artikel 15 (4)	Widerstandsfähigkeit gegenüber Fehlern, Störungen oder Unstimmigkeiten, die innerhalb des Systems oder der Umgebung, in der das System betrieben wird, auftreten können.
Technisch organisatorische Maßnahmen (TOM)	Artikel 15 (4)	Es sind technische und organisatorische Maßnahmen zu treffen, die Robustheit z.B. durch technische Redundanzlösungen erreichen, die Backup- oder Fail-Safe-Pläne umfassen können.
Kontinuierliches Lernen	Artikel 15 (4)	KI-Systeme, die nach dem Inverkehrbringen oder der Inbetriebnahme weiter lernen, sind so zu entwickeln, dass das Risiko, dass möglicherweise verzerrte Ergebnisse den Input für künftige Operationen beeinflussen (Rückkopplungsschleifen), beseitigt oder so weit wie möglich verringert wird, und dass sichergestellt wird, dass solche Rückkopplungsschleifen durch geeignete Abhilfemaßnahmen angemessen berücksichtigt werden.
Cybersicherheit - Widerstandsfähigkeit	Artikel 15 (5)	KI-Systeme müssen gegen Versuche unbefugter Dritter, ihre Nutzung, ihre Ergebnisse oder ihre Leistung durch Ausnutzung von Systemschwachstellen zu verändern, widerstandsfähig sein.
Cybersicherheit - technische Lösungen	Artikel 15 (5)	Die technischen Lösungen müssen den jeweiligen Umständen und Risiken angemessen sein. Sie umfassen Lösungen zur Behebung von KI-spezifischen Schwachstellen, gegebenenfalls Maßnahmen zur Verhinderung, Erkennung, Reaktion, Behebung und Kontrolle von Angriffen, die darauf abzielen, den Trainingsdatensatz oder bereits trainierte Komponenten zu manipulieren, Eingaben, die das KI-Modell zu einem Fehler veranlassen sollen, sowie Angriffe auf die Vertraulichkeit oder Modellfehler.

KI-Systeme mit hohem Risiko müssen gegen Versuche unbefugter Dritter resistent sein, ihre Nutzung, Ergebnisse oder Leistung durch Ausnutzung von Systemschwachstellen zu verändern. Die Anforderungen der KI-VO in Bezug auf Robustheit und Cybersicherheit für KI-Systeme mit hohem Risiko sind in Tabelle 2.1 aufgelistet.

Die Lösungen, die zur Gewährleistung der Cybersicherheit von KI-Systemen mit hohem Risiko eingesetzt werden, müssen den jeweiligen Umständen und Risiken angemessen sein. Zu diesem Zweck müssen technische und organisatorische Maßnahmen getroffen werden. Wir werden uns nur dem Themenkomplex Cybersecurity zuwenden und nicht auf die Robustheit von KI-Systemen eingehen.

2.2 Unterstützende Normen

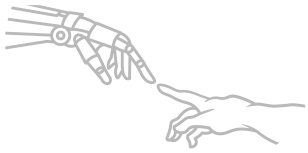
Im Bereich der Cybersicherheit gibt es eine Vielzahl von Normen und Standards, deren Ziel es ist, die Sicherheit von Systemen, Netzwerken und Daten vor potenziellen Bedrohungen zu gewährleisten. Zu den allgemein anerkannten Standards gehört insbesondere die ISO/IEC 27000er Reihe, die sich mit Informationssicherheits-Managementsystemen (ISMS) befasst.

Darüber hinaus gibt es spezielle Standards wie den BSI IT-Grundschutzkatalog des Bundesamtes für Sicherheit in der Informationstechnik (BSI), der Best Practices und Maßnahmen zur Cyber-Sicherheit bereitstellt. Insbesondere bietet das BSI detaillierte Informationen zu KI-Systemen [2].

[2]: BSI (2025), *Künstliche Intelligenz*

Für die Cyber-Sicherheit in kritischen Infrastrukturen ist ebenfalls die ISA/IEC 62443 relevant, die Sicherheitsanforderungen für industrielle Automatisierungs- und Steuerungssysteme beschreibt und auch im Kontext kritischer Infrastruktur Anwendung findet. Die ISA/IEC 62443 besteht aus vier Hauptabschnitten:

- **ISA/IEC 62443-1** ist der allgemeine Abschnitt, der sich mit Terminologie, Modellen und Metriken befasst
- **ISA/IEC 62443-2** befasst sich mit Richtlinien und Verfahren für Sicherheitsmanagement, Implementierungsleitfäden, Patch-Management sowie Installation und Wartung



- **ISA/IEC 62443-3** befasst sich mit der Sicherheit aus der Systemperspektive, d.h. Sicherheitstechnologien, Sicherheitsstufen für Zonen und Netze bzw. Systemsicherheitsanforderungen.
- **ISA/IEC 62443-4** befasst sich mit der Sicherheit der Komponenten, aus denen die in ISA/IEC-62443-3 behandelten Systeme bestehen.

Ergänzend dazu existiert der NIST Cybersecurity Framework (CSF) des National Institute of Standards and Technology, der insbesondere in den USA weit verbreitet ist und einen risikobasierten Ansatz zur Cybersicherheit verfolgt. Das NIST veröffentlicht ebenfalls spezifische Leitlinien zur Vertrauenswürdigkeit von KI, wie den NIST AI Risk Management Framework (AI RMF), der ein strukturiertes Vorgehen zur Identifizierung und Minderung von Risiken in KI-Systemen bietet.

Auf europäischer Ebene hat die ENISA (European Union Agency for Cybersecurity) einen Threat Landscape 2024 Report veröffentlicht mit ähnlichen Risiken und Mitigationen.

Im Bereich der KI befinden sich spezifische Normen noch in der Entwicklungsphase, doch bereits jetzt existieren einige relevante Standards und Leitlinien. Die in der Erarbeitung befindliche ISO/IEC 23894 zielt beispielsweise auf die Festlegung von Sicherheitsanforderungen für KI-Systeme ab. Darüber hinaus bietet der in einer Draft Version vorliegende ISO/IEC DIS 27090 Organisationen einen Leitfaden für den Umgang mit Sicherheitsbedrohungen und Fehlern, die speziell für KI-Systeme gelten. Der Standard soll spezifische Sicherheitsbedrohungen für KI-Systeme und ihre Folgen während ihres gesamten Lebenszyklus untersuchen, und Beschreibungen, wie man solche Bedrohungen erkennen und abschwächen kann, zur Verfügung stellen.

Auch die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) hat mit den „OECD AI Principles“ ethische und sicherheitsrelevante Leitlinien für den Einsatz von KI entwickelt.

Darüber hinaus erarbeitet die ISO/IEC JTC 1/SC 42 eine Reihe von Standards zur KI, die Themen wie Transparenz, Robustheit und Sicherheit abdecken.

Weitere Standards sind ETSI DGR SAI 002 und ETSI GR SAI 002. Hierbei handelt es sich um Sicherheits- und Funktionsstandards für die Sicherheit von KI (Securing Artificial Intelligence (SAI)).

2.3 Sicherheitsziele

Betrachtet man die Sicherheit von KI-Systemen, so gibt es neben den klassischen Sicherheitsaspekten auch KI-spezifische Sicherheitsherausforderungen.

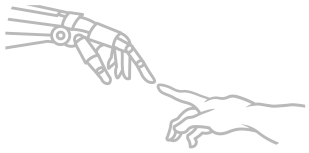
Sicherheitsherausforderungen, die auch für KI-Systeme gelten, sind z.B. die Manipulation von Daten auf dem Weg von der Datenerhebung bis zur Datenbank oder die Bereitstellung von bereits kontaminierten Daten im Internet, z.B. Open Data. Darüber hinaus kann es zu unbefugten Manipulationen an der Datenbank oder zum Diebstahl des Modells kommen. Um personenbezogene Daten zu schützen, Missbrauch zu verhindern und Vertraulichkeit zu gewährleisten, sollten zudem Sicherheitsmechanismen wie Verschlüsselung, Protokollierung von Datenzugriffen und Zugriffskontrollen implementiert werden. Sicherheitsrisiken können natürlich auch in den verwendeten Bibliotheken bestehen. Es stellt sich auch die Frage, wie mit Sicherheitslücken in vortrainierten Modellen, wie z.B. GPAI, umgegangen werden kann und wer dafür verantwortlich ist.

NIST [3] beschreibt einen Überblick über aktuelle Angriffe auf KI-Systeme. Die folgenden Abschnitte basieren auf diesem Dokument. NIST fokussiert zunächst auf Vertraulichkeit, Integrität und Verfügbarkeit. Diese Sicherheitsziele werden in drei Kategorien von Angriffen unterteilt.

[3]: NIST (2023), *Adversarial Machine Learning*

Verletzung der Vertraulichkeit / Privatsphäre (Privacy Compromise)

Angreifer versuchen, Informationen entweder über die Trainingsdaten (Data Privacy) oder über das Modell selbst (Model Privacy) zu erlangen. Mögliche Ziele sind



Datenrekonstruktion, Membership Inference Attacks (Überprüfung, ob bestimmte Daten im Training verwendet wurden) oder Modellextraktion (Extraktion von Modellinformationen).

Integritätsverletzungen (Integrity Violations)

Diese Angriffe zielen darauf ab, falsche Vorhersagen eines Modells zu verursachen. Dazu gehören beispielsweise die Manipulation von Testdaten, um das Modell zu täuschen, sowie verschiedene Arten von Vergiftungsangriffen und Backdoor-Angriffe (Hintertür zur Manipulation von Training und Tests) und Modellvergiftung.

Angriffe auf die Verfügbarkeit (Availability Breakdown)

Hier versucht der Angreifer, die Performance eines Modells im laufenden Betrieb zu beeinträchtigen. Dies kann beispielsweise durch Energy Latency Attacks (Nutzung von Abfragen zur Beeinträchtigung der Effizienz) geschehen.

2.4 Arten von Angriffen

Darüber hinaus kann unterschieden werden, zu welchem Zeitpunkt in der Entwicklung und im Betrieb Angriffe stattfinden können, die die Vertraulichkeit, Integrität und Verfügbarkeit verletzen.

Angriffe während der Datenerhebung

Angriffe während der Datenerhebung bei KI-Systemen zielen darauf ab, fehlerhafte, manipulierte oder schädliche Daten in den Trainingsdatensatz einzuschleusen. Solche Data Poisoning-Angriffe können das Verhalten des Modells gezielt beeinflussen oder seine Leistung verschlechtern. Dadurch wird die Integrität und Verlässlichkeit der KI bereits vor dem eigentlichen Training untergraben.

Angriffe in der Trainingszeit (Poisoning Attacks)

Ein Angreifer kann während der Trainingsphase Trainingsdaten oder Modellparameter manipulieren. Bei Data-Poisoning-Angriffen kontrolliert der Angreifer (Teile der) Trainingsdaten, während bei Modell-Poisoning-Angriffen die Modellparameter verändert werden.

Angriffe während der Testphase

In der Testphase können Angreifer durch gezielte Manipulation von Testdaten die Leistungsbewertung des

Modells verfälschen (z.B. durch Test Data Poisoning). So kann ein unsicheres oder ineffektives Modell fälschlich als zuverlässig eingestuft werden. Zudem könnten vertrauliche Testdaten abgegriffen oder ungewollt offengelegt werden.

Angriffe während des Deployments (Bereitstellung)

Während der Bereitstellung können Angreifer Sicherheitslücken in der Modellintegration oder -konfiguration ausnutzen, um unbefugten Zugriff zu erlangen. Manipulationen in dieser Phase können dazu führen, dass das Modell falsch reagiert oder heimlich Daten preisgibt.

Angriffe während des Betriebs (Inferenzphase)

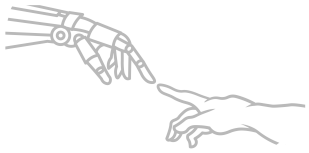
Im laufenden Betrieb sind KI-Systeme anfällig für Adversarial Attacks, bei denen gezielt manipulierte Eingaben kleine Veränderungen verursachen, die jedoch große Fehlschlüsse beim Modell auslösen. Auch Model Inversion kann auftreten, bei dem Angreifer versuchen, Trainingsdaten aus dem Modellverhalten rückzuschließen. Durch Abfragezugriff auf das Modell können Angreifer sensible Informationen über die Trainingsdaten oder das Modell selbst ableiten, z. B. durch Mitgliedschaftsinferenz oder Datenrekonstruktion.

Angriffe während des Decommissionings (Außerbetriebnahme)

Beim Abschalten oder Weitergeben von KI-Systemen besteht das Risiko, dass sensible Daten oder Modellinformationen unzureichend gelöscht werden. Unzureichend gesicherte Altmodelle können von Dritten reaktiviert und missbraucht werden.

Hierzu kann ein Angreifer verschiedene Fähigkeiten einsetzen, um seine Ziele zu erreichen:

- **Kontrolle über die Trainingsdaten:** Der Angreifer kann einen Teil der Trainingsdaten manipulieren, indem er Beispiele einfügt oder verändert. Dies wird bei Data-Poisoning-Angriffen wie Verfügbarkeitsangriffen, gezielten Angriffen oder Backdoor-Poisoning-Angriffen verwendet.
- **Eingeschränkte Kontrolle über Labels:** Bei Clean-Label-Poisoning-Angriffen hat der Angreifer keine Kontrolle über die Labels der Trainingsdaten. Bei regulären Poisoning-Angriffen kann der Angreifer jedoch die Labels manipulieren.



- **Kontrolle über das Modell:** Der Angreifer kann die Modellparameter manipulieren, z.B. durch das Einfügen eines Trojaner-Triggers oder durch das Versenden von bösartigen lokalen Modell-Updates beim föderierten Lernen (Federated Learning).
- **Kontrolle über Testdaten:** Der Angreifer kann während der Ausführung des Modells Störungen in die Testdaten einführen. Dies geschieht bei Umgehungsangriffen, um feindliche Beispiele zu erzeugen, oder bei Backdoor-Poisoning-Angriffen.
- **Kontrolle über den Quellcode:** Der Angreifer kann den Quellcode eines Machine-Learning-Algorithmus verändern, z.B. durch Modifikation des Zufallszahlengenerators oder durch Einfügen von bösartigem Code in Drittanbieter-Bibliotheken, die oft Open-Source sind.
- **Abfragezugriff:** In Fällen, in denen das KI-Modell von einem Cloud-Anbieter verwaltet wird (z.B. bei Machine Learning as a Service - MLaaS), kann der Angreifer Anfragen an das Modell stellen und Vorhersagen (wie Labels oder Modellkonfidenzen) erhalten. Diese Fähigkeit wird bei Black-Box-Angriffen, Latenzangriffen und allen Angriffen auf die Privatsphäre genutzt.

Eine weitere Dimension zur Klassifizierung von Angriffen ist das Wissen des Angreifers über das KI-System. Es gibt drei Hauptarten von Angriffen:

White-Box-Angriffe

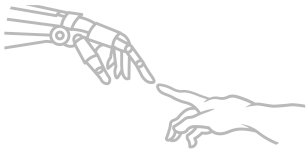
In diesem Fall hat der Angreifer vollständiges Wissen über das KI-System, einschließlich der Trainingsdaten, der Modellarchitektur und der Hyperparameter des Modells.

Black-Box-Angriffe

Der Angreifer hat nur minimales Wissen über das KI-System und nur Abfragezugriff auf das Modell, d.h. das Modell kann nur über öffentlich zugängliche Schnittstellen angegriffen werden

Gray-Box-Angriffe

Diese Angriffe liegen zwischen White-Box- und Black-Box-Angriffen. Der Angreifer hat teilweise Informationen über das KI-System – etwa über die Modellarchitektur oder über Teile der Trainingsdaten – aber nicht



den vollständigen Zugriff wie bei einem White-Box-Angriff.

Einige dieser Angriffe können durch robuste Trainingsmethoden wie Adversarial Training erschwert werden, bei denen das Modell gezielt mit manipulierten (adversarialen) Beispielen trainiert wird, um seine Widerstandsfähigkeit gegenüber solchen Eingaben zu erhöhen. Gerade im Kontext von generativer KI sollten nicht nur Use Cases betrachtet werden, die beschreiben, wie das System „richtig“ genutzt wird, sondern auch Misuse Cases und Abuse Cases definiert werden, die beschreiben, wie das System manipuliert werden kann, um andere Ergebnisse zu berechnen, z.B. die Abfrage von Geheimnissen.

3 Cybersicherheit im Produktlebenszyklus

Dieses Kapitel befasst sich mit den Sicherheitsrisiken, die bei der Entwicklung und dem Betrieb von KI-Systemen auftreten können, und untersucht Technologien, die eingesetzt werden können, um die Entwicklung, den Betrieb und die Stilllegungsphase von KI-Systemen sicherer zu gestalten.

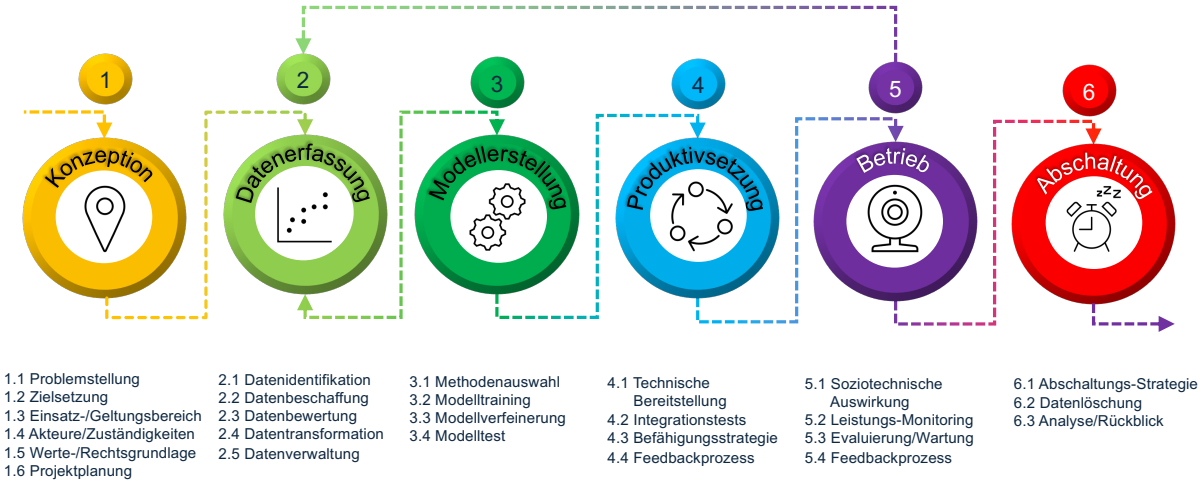


Abbildung 3.1: Der KI-Produktlebenszyklus mit seinen sechs Phasen.

Zur Strukturierung orientieren wir uns an den in Abbildung 3.1 skizzierten Produktlebenszyklus (PLZ) für KI-Systeme. Zur Operationalisierung der Anforderungen aus der KI-VO sind unterschiedliche technische und organisatorische Maßnahmen in den jeweiligen Phasen erforderlich.

3.1 Konzeption

In der Konzeption eines KI-Systems werden die sicherheitsrelevanten Grundlagen gelegt. So ist es zunächst notwendig, rechtliche Anforderungen, welche die KI-VO ergänzen, in der initialen Phase zu integrieren.

Neben Security-spezifischen (z.B. ISO 27001) und KI-spezifischen Standards (wie beispielsweise ISO 27090), sind auch die Datenschutz-Grundverordnung (DSGVO) [4] sowie sektorspezifische Standards zu prüfen, etwa MDR in der Medizin, NIS-2 für Netz- und Informationssysteme oder der Cyber Resilience Act mit Regeln zur Cybersicherheit von Produkten mit digitalen Elementen. Ziel ist es, Risiken frühzeitig zu identifizieren und Maßnahmen zu planen, die Sicherheit, Vertraulichkeit, Integrität und Verfügbarkeit systematisch adressieren. Grundlage hierfür sind, wie in Abschnitt 2.2 beschrieben, etablierte Standards wie etwa ISO/IEC 27001 [5]. Ergänzend ist auch die Nichtabstreitbarkeit sicherzustellen, also die Möglichkeit, Handlungen oder Ereignisse eindeutig bestimmten Akteuren zuzuordnen.

Frühzeitige Bedrohungsanalyse

Bereits in dieser frühen Phase sollte eine Bedrohungsanalyse durchgeführt werden. Dabei sind potenzielle Angriffsvektoren – etwa durch Datenmanipulation, Modellinversion oder adversarielle Angriffe – systematisch zu erfassen. Zur Umsetzung empfiehlt sich die Verwendung gängiger Methoden wie *STRIDE*, einem Modell zur Kategorisierung von Bedrohungen anhand von sechs Typen (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service und Elevation of Privilege), oder *Attack Trees*, einer hierarchischen Darstellung möglicher Angriffspfade, bei der ein übergeordnetes Ziel in Teilziele und konkrete Angriffsschritte zerlegt wird. Abbildung 3.2 illustriert beispielhaft einen solchen Angriffsbaum auf das Modell eines KI-Systems, in dem verschiedene Manipulationsmöglichkeiten systematisch aufgegliedert sind.

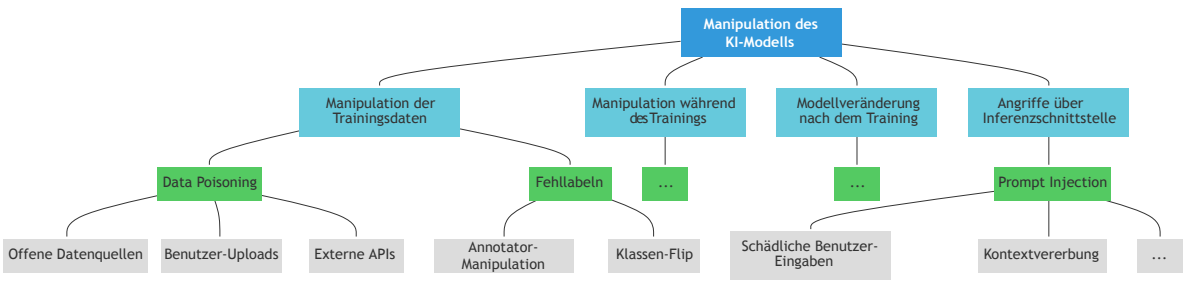
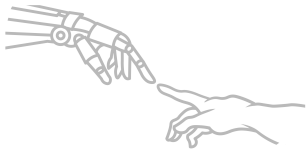


Abbildung 3.2: Minimalbeispiel eines Attack Trees auf das Modell eines KI-Systems.



Frühzeitige Risikoanalyse

Ergänzend ist eine Risikobewertung durchzuführen, bei der für jede identifizierte Bedrohung die Eintrittswahrscheinlichkeit und das potenzielle Schadensausmaß geschätzt werden, um die Relevanz der Bedrohung quantifizierbar zu machen. Die daraus abgeleiteten Maßnahmen sind zu priorisieren und dokumentiert in das Sicherheitskonzept aufzunehmen.

Konzept für Datenherkunft, -qualität und -sicherheit

Zur Absicherung der Datenverarbeitung ist ein Konzept für Datenherkunft, -qualität und -sicherheit zu entwickeln. Dies umfasst unter anderem Anforderungen an Datenquellen, Prüfmechanismen zur Erkennung von Anomalien und technische Verfahren zur Datenverarbeitung wie Anonymisierung und Pseudonymisierung. Die Auswahl dieser Verfahren sollte im Hinblick auf den Anwendungsfall, die Schutzbedarfsanalyse und die Rückverfolgbarkeit erfolgen.

Technische und organisatorische Schutzmaßnahmen

Im Hinblick auf die spätere Entwicklung sind technische und organisatorische Schutzmaßnahmen vorzubereiten. Dazu gehört die Einrichtung einer sicheren Entwicklungsumgebung mit Zugriffskontrollen, Versionierung und Protokollierung von Änderungen an Daten und Modellen. Die Protokollierung sollte dabei möglichst so granular erfolgen, dass sicherheitsrelevante Ereignisse nachvollziehbar sind, ohne unnötig Speicherressourcen zu binden.

Security by Design

Ein weiterer Aspekt ist die Integration des Prinzips Security by Design. Dies erfordert, dass Sicherheitsanforderungen nicht nachträglich ergänzt, sondern von Beginn an in Architektur und Design berücksichtigt werden. Das Entwicklungsteam ist dafür zu sensibilisieren, etwa durch gezielte Schulungen zu sicherheitsrelevanten Aspekten von KI-Systemen, typischen Bedrohungsszenarien und Best Practices zur Absicherung von Trainingsdaten und Modellen.

Supply Chain Security

Die frühzeitige Berücksichtigung von Supply Chain Security in der Konzeptionsphase ist entscheidend, um Sicherheitsrisiken entlang der gesamten Lieferkette systematisch zu identifizieren und zu minimieren. Bereits bei der Planung neuer Produkte, Systeme oder Prozesse

lassen sich durch geeignete Maßnahmen – wie Risikoanalysen, Lieferantenbewertungen und Sicherheitsanforderungen – potenzielle Schwachstellen proaktiv adressieren. Dies ermöglicht nicht nur eine widerstandsfähigere und vertrauenswürdigere Lieferkette, sondern verhindert auch kostenintensive Nachbesserungen und reduziert die Anfälligkeit für Angriffe, Manipulationen oder Ausfälle in späteren Projektphasen.

Konzept für Wirksamkeitsprüfung

Auch ist frühzeitig festzulegen, wie die Wirksamkeit der geplanten Maßnahmen überprüft wird. Dazu gehören regelmäßige Sicherheitsreviews, Codeanalysen, Penetrationstests und Audits. Diese Prüfmechanismen sind in den Projektplan zu integrieren, um eine kontinuierliche Validierung der Sicherheitsmaßnahmen entlang des gesamten Lebenszyklus zu ermöglichen.

3.2 Datenerfassung

Die Phase der Datenerfassung,- -verarbeitung und -nutzung bildet das Fundament jedes KI-Systems – sowohl funktional als auch sicherheitstechnisch. Da KI-Modelle typischerweise auf umfangreichen Datenmengen basieren, ergeben sich in dieser Phase erhebliche Anforderungen an Datenschutz, Vertraulichkeit, Integrität und Verfügbarkeit. Um einerseits gesetzlichen Verpflichtungen – etwa im Rahmen der DSGVO – und andererseits den Anforderungen an Cybersicherheit und Modellzuverlässigkeit gerecht zu werden, bedarf es eines systematischen, technisch unterstützten Umgangs mit den Daten. Tabelle 3.3 gibt eine Übersicht über sicherheitsrelevante Risiken in verschiedenen Phasen der Datenverarbeitung in KI-Systemen. Die Matrix zeigt, in welcher Phase ein Risiko besonders relevant ist, dargestellt durch die Größe der Kreise. Die Farbgebung der Kreise verweist auf jeweils betroffene Schutzziele. So wird auf einen Blick erkennbar, wann und wo im Datenlebenszyklus bestimmte Bedrohungen besonders beachtet werden sollten. Im Folgenden wird aufgezeigt,

wie wesentliche Schutzmaßnahmen während der Datenverarbeitung konkret umgesetzt werden können.

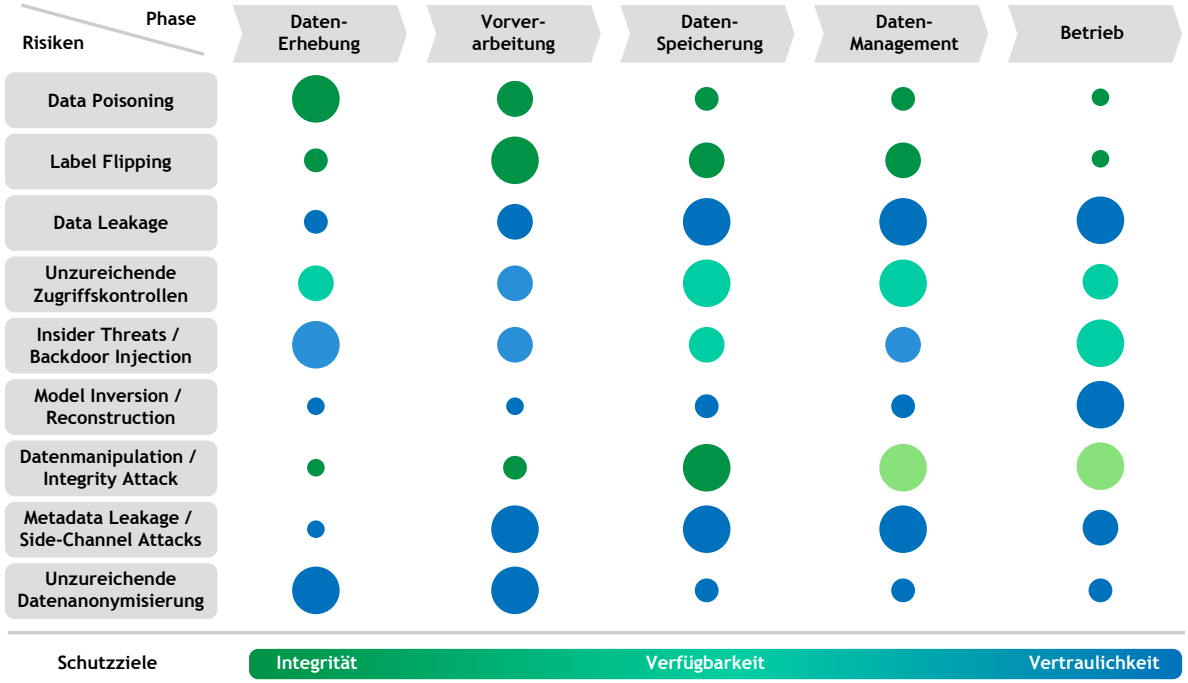


Abbildung 3.3: Übersicht über sicherheitsrelevante Risiken in verschiedenen Phasen der Datenverarbeitung in KI-Systemen.

Rechtssichere Datenerhebung und Datenschutz Ein erster Schritt ist die rechtssichere und transparente Datenerhebung. Die Einhaltung datenschutzrechtlicher Anforderungen beginnt bereits vor der Modellierung, nämlich bei der Auswahl, Erhebung und Verarbeitung der Daten. In der Praxis bedeutet das: Es muss eine klare Rechtsgrundlage für die Verarbeitung vorliegen – etwa eine informierte Einwilligung der betroffenen Personen, ein Vertrag oder ein berechtigtes Interesse. Die betroffenen Personen sind über Art, Zweck und Umfang der Datenverarbeitung transparent zu informieren. Insbesondere bei potenziell risikobehafteten Anwendungen ist zusätzlich eine Datenschutz-Folgenabschätzung gemäß Art. 35 DSGVO [4] durchzuführen.

Datenminimierung

Dabei sollte von Anfang an das Prinzip der Datenminimierung verfolgt werden: Nur die für den jeweiligen Verwendungszweck unbedingt notwendigen Daten soll-

ten erhoben und gespeichert werden. Veraltete oder redundante Daten sind regelmäßig zu löschen oder zu anonymisieren. Zusätzlich sollten Mechanismen zur Überwachung und Protokollierung verdächtiger Aktivitäten etabliert werden, um potenzielle Angriffe frühzeitig zu erkennen.

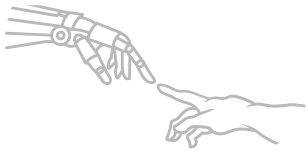
Data Leakage-Effekte

Dabei ist auf sogenannte Data Leakage-Effekte zu achten – also die unbeabsichtigte Einbeziehung von Informationen in das Training, die zum Zeitpunkt der Modellnutzung nicht verfügbar wären. Solche Leckagen können zu überoptimistischen Leistungsmetriken führen und die Aussagekraft des Modells untergraben [6]. Besondere Vorsicht ist in diesem Kontext bei der Aufteilung in Trainings-, Validierungs- und Testdaten geboten: Leckagen zwischen den Datensätzen (Train/Test Leakage) – z.B. durch versehentlich doppelt vorkommende oder falsch vorverarbeitete Daten – können die Modellgüte verfälschen und sollten durch saubere Trennung und auditierbare Pipelines verhindert werden.

Datenqualität und Schutz vor Datenmanipulation
Parallel zur rechtlichen Absicherung ist die Sicherstellung der Datenqualität und -integrität ein zentraler Bestandteil sicherer KI-Entwicklung. Fehlerhafte, unvollständige oder manipulierte Daten können zu schwerwiegenden Verzerrungen oder Fehlfunktionen im Modell führen. So bergen insbesondere Data Poisoning- und Label Flipping-Angriffe erhebliche Risiken, bei denen gezielt fehlerhafte oder falsch etikettierte Daten in das Trainingsset eingeschleust werden. Eine (kontinuierlich) hohe Datenqualität wird durch automatisierte Validierungsprozesse gewährleistet, die etwa Wertebereiche, Formatkonsistenz und Vollständigkeit prüfen (siehe z.B. [7], [8]).

Data Poisoning

Während Data Poisoning-Angriffe bereits allgemein als wesentliches Risiko benannt wurden, unterscheidet der ISO-Standard ISO/IEC DIS 27090 explizit zwischen zwei Arten: gezielten (targeted) und wahllosen (indiscriminate) Datenvergiftungsangriffen. Gezielte Angriffe haben



zum Ziel, spezifische Beispiele im Datensatz absichtlich fehlzuklassifizieren. Dabei werden oft sogenannte Trigger-Merkmale eingeschleust, die ein Modell gezielt in bestimmten Situationen fehlleiten. Im Gegensatz dazu streben wahllose Datenvergiftungsangriffe danach, die generelle Integrität und Zuverlässigkeit des gesamten Trainingsdatensatzes zu beeinträchtigen.

In der Praxis zeigen sich solche Angriffe in vielfältigen Anwendungsbereichen wie etwa bei Spamfiltern, Malware-Erkennungssystemen, automatischer Code-Vervollständigung oder medizinischen Diagnosemodellen. Für die Erkennung solcher Angriffe haben sich fortgeschrittene Methoden bewährt, beispielsweise die Verwendung mehrerer kleiner Modelle („Mikromodelle“) auf getrennten Teilen des Datensatzes mit anschließender Mehrheitsentscheidung. Auch Verfahren wie Trigger-Inversion zur Identifikation versteckter Backdoors oder Aktivierungs-Clustering, um auffällige Muster im Modell zu erkennen, sind effektiv.

Als Gegenmaßnahmen gegen Datenvergiftung bieten sich robuste algorithmische Verfahren an, etwa der Einsatz von Generative Adversarial Networks (GANs), Modell-Ensembles und robuste Optimierungsverfahren beim Training. Ergänzend helfen gezielte Datenerweiterungen (Data Augmentation), sowie Techniken wie *Fine-Pruning* oder *Fine-Tuning*, um die Empfindlichkeit eines Modells gegenüber manipulierten Daten deutlich zu reduzieren.

Technisch-organisatorische Sicherheitsmaßnahmen
Wie grundsätzlich in der Datenerfassungsphase vorgesehen, empfiehlt sich im Hinblick auf die Entwicklung eines sicheren und robusten KI-Systems eine sorgfältige Datenbereinigung etwa durch die Entfernung von Duplikaten, Prüfung auf Inkonsistenzen sowie auf unplausible Ausreißer. Datenquellen sollten bevorzugt verifiziert und – soweit möglich – zertifiziert sein. Zum Schutz vor nachträglicher Manipulation bieten sich technische Maßnahmen wie die Verwendung von digitalen

Signaturen, Hash-Werten (z.B. SHA-256) oder Checksummen an. Auch Synthetic Data Injection lässt sich durch Herkunftsnachweise und automatisierte Konsistenzprüfungen deutlich erschweren. Zudem ist sicherzustellen, dass ebenfalls die Softwarekomponenten, die im Rahmen des Supply Chain Managements zum Einsatz kommen, verifiziert, vertrauenswürdig und zuverlässig sind.

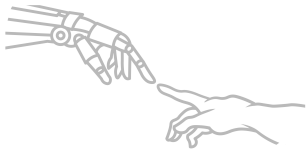
Technische Zugriffssicherheit

Die technische Zugriffssicherheit muss hierbei berücksichtigt werden, indem beispielsweise der Zugriff auf sensible Daten ausschließlich autorisierten Personen gewährt werden darf. Dies kann durch die Implementierung eines Identity and Access Management (IAM) erreicht werden (siehe z.B. NIST [9]), das Rollen- (RBAC) oder attributbasierte Zugriffskontrollsysteme (ABAC) einsetzt (siehe z.B. [10]). Ergänzt werden diese durch Authentifizierungsverfahren wie Single Sign-On (SSO), Multi-Faktor-Authentifizierung (MFA) oder – bei höheren Schutzanforderungen – biometrische Verfahren.

Alle Zugriffe und Änderungen an den Daten sollten protokolliert und in diesem Zusammenhang regelmäßig auf Anomalien überprüft werden. Dies bietet darüber hinaus Schutz vor *Insider Threats* und unbefugter Modifikation, die etwa im Rahmen von *Trojaning* oder *Backdoor Attacks* in die Daten eingeschleust werden könnten. Auf diese Weise wird nicht nur die Vertraulichkeit gewahrt, sondern auch potenzieller Missbrauch frühzeitig erkannt.

Eine vergleichende Betrachtung verschiedener IAM-Methoden kann helfen, geeignete Maßnahmen auszuwählen. Dabei weisen z.B. Passwort-basierte Verfahren eine hohe Anfälligkeit gegenüber Phishing auf, während biometrische Authentifizierung zwar sicher, aber kostenintensiv ist. RBAC erlaubt einfache Verwaltung durch Rollenzuweisung, während ABAC eine feinere Kontrolle ermöglicht, jedoch eine höhere Implementierungskomplexität mit sich bringt.

Zusätzlich zu allgemeinen Zugriffskontrollen ist es sinnvoll, spezifische Zero-Trust-Prinzipien auf Trainingsda-



tensätze und KI-Modelle anzuwenden. Insbesondere das Prinzip *Assume Breach*, bei dem grundsätzlich angenommen wird, dass ein Sicherheitsbruch bereits eingetreten sein könnte, bietet eine wichtige strategische Grundlage. Dies führt dazu, dass ständig aktive und sorgfältig abgestimmte Schutz- und Erkennungsmaßnahmen implementiert sind, die auch kleinste Unregelmäßigkeiten frühzeitig sichtbar machen.

Das ergänzende Prinzip *Least Privilege Access* sollte konsequent angewendet werden, um Zugriffe auf sensible Daten und Modelle streng zu limitieren. Praktisch bedeutet das, Zugriffsrechte nur zeitlich begrenzt (*Just-in-Time*) und nur im minimal notwendigen Umfang (*Just-Enough-Access*) zu vergeben. Solche Maßnahmen helfen, die Angriffsfläche deutlich zu reduzieren und sowohl interne als auch externe Risiken zu minimieren.

Verschlüsselung

Ein weiteres wesentliches Sicherheitsprinzip ist die durchgehende Verschlüsselung der Daten. Dabei ist zwischen der Verschlüsselung bei der Übertragung (z.B. mittels TLS) und der Verschlüsselung im Ruhezustand (z.B. AES-256 für gespeicherte Daten) zu unterscheiden. Für die sichere Verwaltung kryptographischer Schlüssel bietet sich die Einrichtung einer Public Key Infrastructure (PKI) an. Es ist zwischen symmetrischer Verschlüsselung (z.B. AES-256) und asymmetrischer Verschlüsselung (z.B. RSA, ECC) zu differenzieren.

In sensiblen Szenarien können zudem fortgeschrittene kryptographische Verfahren wie homomorphe Verschlüsselung oder hardwaregestützte sichere Ausführungsumgebungen (z.B. Intel SGX) in Betracht gezogen werden, um selbst während der Verarbeitung keine Klartextdaten preiszugeben [11]. Diese Maßnahmen helfen insbesondere gegen Model Inversion- oder Data Reconstruction-Angriffe, bei denen aus Modellzugriffen Rückschlüsse auf Originaldaten gezogen werden sollen. Mit Blick auf die Zukunft ist ebenfalls der Einsatz quantensicherer Verschlüsselungsverfahren zu erwägen.

Datenschutzkonforme Tarnung

Zusätzlich zur Absicherung gegen externe Bedrohungen spielt die datenschutzkonforme Tarnung personenbezogener Informationen eine wichtige Rolle. Techniken wie Anonymisierung und Pseudonymisierung helfen dabei, die Rückführbarkeit auf Einzelpersonen zu minimieren. Für Testszenarien oder nicht-produktive Umgebungen eignen sich Verfahren wie *Tokenisierung*, *Maskierung*, *Daten-Shuffling* oder das gezielte Hinzufügen von Rauschen. Methoden wie *Private Aggregation of Teacher Ensembles (PATE)* oder *Differential Privacy* können je nach Anwendungsbereich genutzt werden, um den Schutz individueller Daten selbst während des Trainings sicherzustellen und zugleich *Membership Inference*- und *Model Inversion*-Risiken zu minimieren. Ergänzend kann Datenredaktion eingesetzt werden, z.B. durch das Zensieren sensibler Bereiche auf Bildern oder Schwärzen von Textpassagen in Dokumenten.

Im Hinblick auf konkrete Angriffsvektoren, wie etwa Data-Poisoning-Angriffe, ist eine Kombination aus technischen und organisatorischen Maßnahmen erforderlich. Um zu verhindern, dass manipulierte Daten in das Training gelangen, sollten die oben erwähnten automatisierten Prüfprozesse mit Anomalieerkennung sowie Datenherkunftsnachweise (Data Provenance) eingesetzt werden.

Nicht zuletzt muss die technische Infrastruktur berücksichtigt werden: Die Speicherung und Verarbeitung sensibler Daten sollte im Idealfall ausschließlich in zertifizierten Rechenzentren erfolgen, etwa nach ISO 27001. Regelmäßige Backups, Zugriffsaudits sowie die kontinuierliche Schulung und Sensibilisierung der Entwicklerteams im Umgang mit Datenschutz und Cybersicherheit tragen dazu bei, Sicherheitsstandards langfristig aufrechtzuerhalten. Die Integration von Monitoring- und Logging-Funktionen unterstützt zusätzlich bei der frühzeitigen Erkennung von Angriffen.

Einen Überblick über weitere relevante Angriffsvektoren sowie praxisbewährte Schutzmaßnahmen in der Datenerfassungsphase von KI-Systemen liefert Tabelle 3.1.

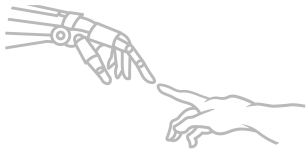


Tabelle 3.1: Auswahl häufiger Angriffsvektoren mit entsprechenden Schutzmaßnahmen in der Datenerfassungsphase von KI-Systemen

Risiko / Angriffsvektor	Beschreibung	Absicherungsmaßnahmen
Data Poisoning	Einschleusen manipulierter oder falsch gelabelter Trainingsdaten	▶ Validierung der Daten (z.B. Wertebereiche, Formate) ▶ Anomalieerkennung ▶ Herkunftsnachweise (Data Provenance) ▶ Nutzung zertifizierter Datenquellen
Label Flipping	Zielgerichtete Umkehrung von Datenlabels zur Modellmanipulation	▶ Automatisierte Label-Konsistenzprüfung ▶ Redundante Quellenvalidierung ▶ Manuelle Stichprobenprüfung
Data Leakage	Verwendung von Informationen im Training, die im Echtbetrieb nicht verfügbar wären	▶ Feature-Auswahl mit Blick auf Echtzeitverfügbarkeit ▶ Simulationsprüfung ▶ Auditierbare Vorverarbeitungs-Pipelines (Daten Split) ▶ Feature Leakage Detection Tools ▶ Hash-Vergleiche zur Duplikaterkennung
Unzureichende Zugriffskontrollen	Unbefugter Zugriff auf sensible Daten durch mangelhaftes Identitätsmanagement	▶ Implementierung von IAM-Systemen (z.B. RBAC, ABAC) ▶ MFA, SSO, biometrische Authentifizierung ▶ Protokollierung und regelmäßiges Audit
Insider Threats Backdoor Injection	Modifikation durch interne Akteure oder schadhafte Komponenten	▶ Logging aller Änderungen ▶ Zugriffsbeschränkung nach Least-Privilege-Prinzip ▶ Verifizierung von Softwarekomponenten (Supply Chain Security)
Model Inversion Reconstruction	Rekonstruktion sensibler Trainingsdaten aus Modellausgaben	▶ Differential Privacy ▶ Homomorphe Verschlüsselung ▶ Techniken wie PATE ▶ Minimierung der Modellantworten
Datenmanipulation Integrity Attack	Veränderung von Daten während Übertragung oder Speicherung	▶ Verschlüsselte Kommunikationswege, z.B. TLS ▶ Verschlüsselung für gespeicherte Daten, z.B. AES-256 ▶ Digitale Signaturen / Hashes / Checksummen ▶ Nutzung von PKI zur Schlüsselverwaltung
Metadata Leakage Side-Channel Attacks	Rückschlüsse über Metadaten (z.B. Zugriffszeiten, Speicherorte)	▶ Minimierung sensibler Metadaten ▶ Abschirmung kritischer Prozesse ▶ Nutzung vertrauenswürdiger Hardware (z.B. SGX)
Unzureichende Datenanonymisierung	Ungewollte Rückverfolgbarkeit auf Trainingsdaten	▶ Kombination mehrerer Techniken (Anonymisierung, Tokenisierung, Shuffling, Rauschen) ▶ Trennung von Identifikations- und Nutzungsinstanzen

3.3 Modellerstellung

Die Modellerstellungsphase folgt im Lebenszyklus von KI-Systemen nach der Datenerfassungsphase und deckt nicht nur im Hinblick auf Leistung und Funktionalität des KI-Systems einen entscheidenden Bereich ab, sondern ist auch im Hinblick auf die Cybersicherheit relevant. In dieser Phase wird das Verhalten des KI-Systems trainiert, da Architekturwahl, Trainingsmetho-

den und Modellparameter langfristige Auswirkungen auf Robustheit, Transparenz und Vertrauenswürdigkeit haben. Hier entstehen neue Angriffsflächen - von gezielten Manipulationen der Modellstruktur bis hin zu versteckten Schwachstellen in komplexen neuronalen Netzwerken.

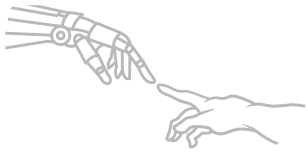
Wie bereits in der Datenerfassungsphase (Abschnitt 3.2) ist auch hier eine systematische, sicherheitsbewusste Entwicklung erforderlich. Nur so lässt sich gewährleisten, dass die Schutzziele auch im operativen Einsatz erhalten bleiben. Die folgenden Abschnitte beleuchten konkrete Risiken und Schutzmaßnahmen, die während der Modellerstellung berücksichtigt werden sollten, um KI-Systeme widerstandsfähig gegenüber aktuellen und zukünftigen Bedrohungen zu gestalten.

Risiken spezifischer Modellarchitekturen Bestimmte Modellarchitekturen weisen eine höhere Anfälligkeit gegenüber Angriffen auf, insbesondere durch sogenannte adversariale Beispiele (*Adversarial Examples*). Diese Angriffe manipulieren Eingabedaten gezielt, um Modelle zu Fehlklassifikationen zu verleiten, während die Veränderungen für menschliche Beobachter oft kaum wahrnehmbar bleiben.

Tiefe neuronale Netzwerke

Vor allem tiefe neuronale Netzwerke (*Deep Neural Networks*, *DNNs*), wie *Convolutional Neural Networks* (*CNNs*), sind betroffen, da ihre Entscheidungsgrenzen durch kleine Veränderungen leicht beeinflussbar sind. Beispielsweise könnten leicht modifizierte Verkehrsschilder fälschlich interpretiert werden, was gravierende Folgen in sicherheitskritischen Anwendungen haben kann.

Ein typisches Beispiel ist die gezielte Manipulation eines Stoppschildes durch kaum sichtbare Aufkleber oder kleine Änderungen, welche dazu führen können, dass autonome Fahrzeuge das Schild nicht erkennen oder es falsch interpretieren. Ebenso anfällig sind Gesichtserkennungssysteme, bei denen minimale Änderungen in der Eingabe, die für Menschen unsichtbar bleiben, zu fehlerhaften Identifizierungen führen können.



Recurrent Neural Networks

Auch Recurrent Neural Networks (*RNNs*), die häufig für Text- und Sprachverarbeitung genutzt werden, zeigen eine hohe Sensitivität gegenüber adversarialen Angriffen. Bereits kleine, geschickt eingefügte Wörter oder Veränderungen in der Satzstruktur können die Vorhersagequalität deutlich beeinträchtigen. Dies betrifft beispielsweise automatisierte Übersetzungssysteme oder sentimentbasierte Anwendungen, die durch subtile Eingabeänderungen massiv fehlgeleitet werden können.

Transformer-Modelle

Transformer-Modelle, die insbesondere im Bereich der natürlichen Sprachverarbeitung (*NLP*) und Bildanalyse weit verbreitet sind, besitzen ebenfalls Angriffspunkte. Ihre Fähigkeit, Kontextinformationen dynamisch zu gewichten und zu kombinieren, kann durch speziell gestaltete Eingaben manipuliert werden. In der Praxis könnten subtile Änderungen in Texten beispielsweise dazu führen, dass Spamfilter getäuscht werden oder Systeme falsche Zusammenfassungen generieren.

Sicherere Alternativen

Als sicherere Alternativen gelten Modelle, die explizit auf Robustheit optimiert sind, etwa robuste lineare Modelle, *Support Vector Machines* (*SVMs*) oder spezialisierte neuronale Netzwerke für sicherheitskritische Anwendungen. Diese Modelle besitzen häufig klar definierte Entscheidungsgrenzen oder integrieren von Beginn an Mechanismen, welche die Empfindlichkeit gegenüber adversarialen Veränderungen reduzieren.

Adversariales Training

Um diese architekturbedingten Risiken zu minimieren, sollten Modelle gezielt durch adversariales Training gehärtet werden. Dabei wird das Modell während des Trainings bereits mit manipulierten Eingaben konfrontiert, um die Robustheit gegenüber solchen Angriffen zu steigern. Weitere effektive Maßnahmen umfassen *Gradient Clipping*, bei dem Gradienten während des Trainingsprozesses bewusst begrenzt werden, um die Anfälligkeit gegenüber kleinen Veränderungen zu reduzieren. *Defensive Distillation* stellt eine weitere Technik dar, bei der das Modell gezielt darauf trainiert wird, weniger empfindlich gegenüber adversarialen Beispielen zu reagieren, indem interne Repräsentationen und Entscheidungsschwellen robust gestaltet werden.

Input-Preprocessing

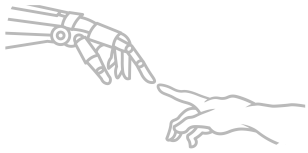
Auch Input-Preprocessing spielt eine entscheidende Rolle bei der Verbesserung der Robustheit von KI-Modellen. Hierbei werden potenziell gefährliche Eingabedaten durch Filterung, Glättung oder Normalisierung modifiziert, um subtile Manipulationen abzuschwächen oder vollständig zu eliminieren.

Formale Verifikation

Ein weiterer Ansatz zur Steigerung der Modellrobustheit sind formale Verifikationstechniken, welche eine mathematisch fundierte Analyse des Modells ermöglichen. Durch solche Verfahren können formale Garantien z.B. über die Robustheit gegenüber adversarialen Angriffen gegeben werden.

Die Kombination dieser Maßnahmen trägt entscheidend dazu bei, die spezifischen Risiken unterschiedlicher Modellarchitekturen zu adressieren und sichere KI-Systeme zu entwickeln. Um diese architekturbedingten Risiken zu minimieren, sollten zusammengefasst folgende Maßnahmen umgesetzt werden:

- **Adversariales Training:** Integrieren Sie bewusst manipulierte Eingaben ins Training, um das Modell frühzeitig auf typische Angriffsformen zu sensibilisieren.
- **Gradient Clipping:** Begrenzen Sie die Gradientenwände im Training, um übermäßige Sensitivität auf kleine Eingabeveränderungen zu vermeiden.
- **Defensive Distillation:** Verwenden Sie Modelle mit robusten Entscheidungsgrenzen, die weniger anfällig für gezielte Manipulationen sind.
- **Input-Preprocessing:** Führen Sie Filter, Normalisierungen oder Störungsunterdrückung auf Eingabedaten durch, bevor diese dem Modell zugeführt werden.
- **Formale Verifikation:** Nutzen Sie mathematische Verfahren zur Absicherung, dass das Modell sich innerhalb definierter Grenzen verhält.
- **Architekturwahl:** Wählen Sie Modellarchitekturen unter Berücksichtigung der konkreten Risikoszenarien (z.B. möglichst interpretierbare Modelle in kritischen Kontexten).



Absicherung der Trainingsprozesse Während der Trainingsphase eines KI-Modells werden die zentralen Parameter und das Verhalten des Systems geprägt. Zugleich stellt diese Phase als Übergang von Daten zur Produktivsetzung einen sensiblen Angriffspunkt dar: Manipulierte Trainingsdaten, fehlerhafte Aggregationsmechanismen oder unzureichend geschützte Trainingsumgebungen können Sicherheitslücken in das KI-Modell einschleusen. Besonders bei verteilten Trainingsansätzen oder der Nutzung externer Datenquellen steigt das Risiko signifikant.

Absicherung der Trainingsumgebung

Bei der Absicherung des Trainingsprozesses besteht ein erster Schritt darin, die Trainingsumgebung selbst abzusichern. Modelle sollten ausschließlich auf gehärteten Systemen trainiert werden – etwa in isolierten Containern, abgeschotteten virtuellen Maschinen oder auf Hardware mit vertrauenswürdigen Ausführungsumgebungen wie TPM oder SGX. So wird verhindert, dass externe oder nicht autorisierte interne Prozesse Zugriff auf sensible Zwischenergebnisse, Parameter oder Speicherbereiche erhalten.

Nachvollziehbarkeit

Auch die Nachvollziehbarkeit spielt eine wichtige Rolle: Jedes Modelltraining sollte vollständig dokumentiert und versioniert werden – inklusive der verwendeten Daten, der Hyperparameter, des Codes sowie aller Modellartefakte. Tools wie MLflow (für Modelle), DVC (für Daten) oder Git (für Code) in Kombination mit Logging-Mechanismen ermöglichen die Wiederherstellung einzelner Trainingsschritte und schaffen die Grundlage für spätere Audits oder Incident-Analysen.

Validierungsmechanismen

Ein besonderes Augenmerk sollte auf die Integrität und Vertrauenswürdigkeit der Trainingsdaten gelegt werden. Bei extern zugeführten oder bei verteilten Lernverfahren eingesammelten Modelle oder Modell-Teile können Angreifer gezielt versuchen, bösartige Muster – sogenannte *Backdoors* – in das Produktiv-Modell einzuschleusen. Ähnlich wie bei den Daten (vgl. Abschnitt 3.2), helfen in Bezug auf das Modell Validierungsmechanismen wie etwa digitale Signaturen, Prüfsummen oder Herkunftsnachweise, aber auch algorithmische Methoden

wie Trigger-Inversion oder Aktivierungsanalysen, um versteckte Manipulationen aufzudecken.

Modellvergiftung

Analog zur Datenvergiftung existiert besonders im Kontext der Modellierung eine prominente Bedrohung: die direkte Modellvergiftung. Hierbei greifen Angreifer nicht über die Trainingsdaten ein, sondern manipulieren gezielt technische Komponenten innerhalb der Modellierungs- oder Entwicklungsumgebung. Dies kann verschiedene Formen annehmen – von der Veränderung des Trainingscodes über die Einschleusung kompromittierter Bibliotheken bis hin zur Modifikation bereits gespeicherter Modellartefakte.

Um einem vollständigen Modellvergiftungsszenario vorzubeugen, empfiehlt sich grundlegend der Einsatz robuster Trainingsverfahren. Dazu zählen z.B. adversariales Training mit gezielt manipulierten Eingaben, um die Widerstandskraft des Modells systematisch zu erhöhen, sowie Methoden wie *Fine-Pruning* oder *Fine-Tuning*, mit denen schadhafte Muster aus einem bereits trainierten Modell entfernt werden können. In verteilten Umgebungen wiederum bieten robuste Aggregationsverfahren wie *Krum*, *Trimmed Mean* oder *Median* eine Möglichkeit, fehlerhafte oder absichtlich schadhafte Client-Modelle zu identifizieren und deren Einfluss zu minimieren.

Ein typisches Szenario ist die Integration eines präparierten Modells in eine Build-Pipeline, das bereits eine Backdoor enthält. Diese wird durch einen simplen Trigger – beispielsweise eine bestimmte Eingabekombination – aktiviert und führt dann zu gezielt falschem Modellverhalten. Besonders in Organisationen, die vortrainierte Modelle von Dritten beziehen oder offene Repositories nutzen, stellt dies ein reales Risiko dar. Eine kleine, unbemerkte Änderung an einem Modell kann dazu führen, dass es bei spezifischen Eingaben abweicht, ohne dass dies während der Validierung auffällt.

Manipulation des Trainingscodes

Ebenso kritisch sind manipulative Eingriffe in den Trainingscode selbst. Beispielsweise kann durch eine scheinbar harmlose Veränderung an einem Loss- oder Preprocessing-Modul gezielt bewirkt werden, dass das Modell

auf bestimmte Muster besonders empfindlich oder blind reagiert. Auch bei automatisierten *MLOps*-Pipelines – in denen Trainingsprozesse regelmäßig ohne menschliches Zutun angestoßen werden – kann ein gezielter Eingriff in die Skripte oder Parameterkonfiguration weitreichende Folgen haben.

Als Schutzmaßnahmen empfiehlt sich ein mehrschichtiger Ansatz, welcher in Abbildung 3.4 zu dargstellt ist.

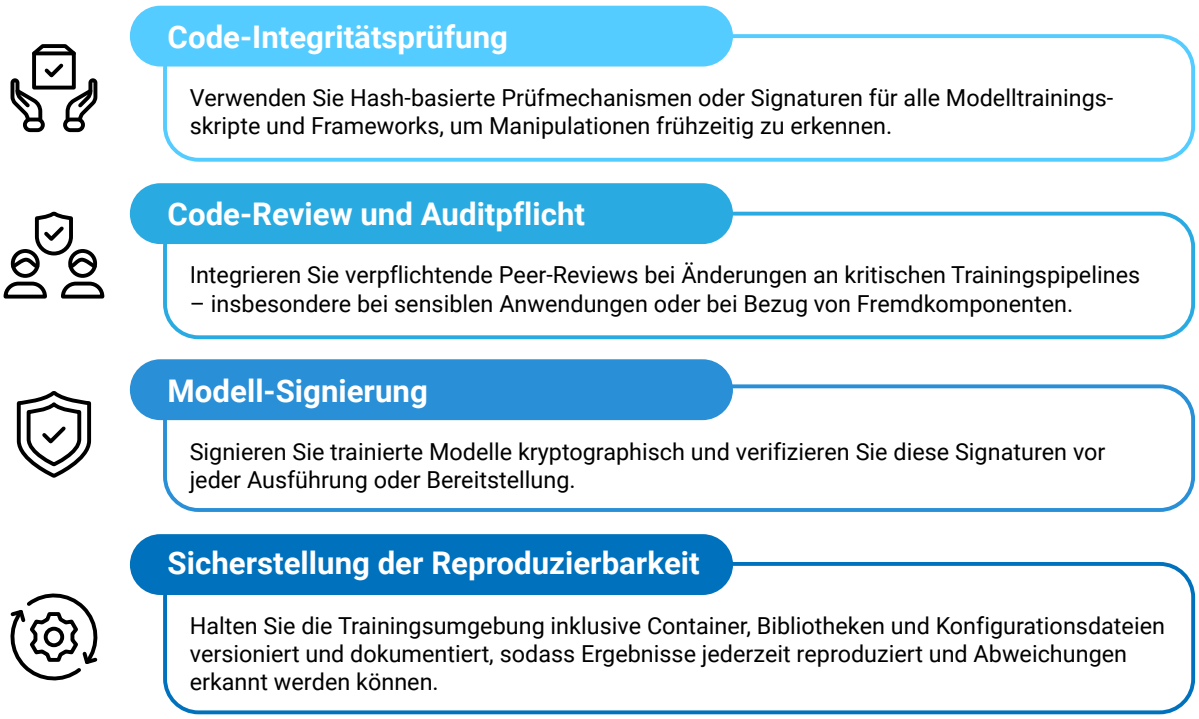


Abbildung 3.4: Abwehrstrategien gegen Model-Poisoning zur Trainings-/Entwicklungszeit: Maßnahmen zur Verhinderung der gezielten Einschleusung manipulativer Funktionalitäten in KI-Modelle durch Angreifer.

Schutz sensibler Informationen

Darüber hinaus sollte in sicherheitskritischen Anwendungen auch der Schutz sensibler Informationen im Training beachtet werden. Obwohl ebenfalls bereits in Abschnitt 3.2 erwähnt, sind Verfahren wie *Differential Privacy* in der Modellierungsphase relevant und können z.B. um *Gradient Clipping* ergänzt werden. Diese Methoden helfen dabei, Rückschlüsse auf einzelne Trainingsbeispiele zu verhindern – ein besonders wichtiger Aspekt beispielsweise im medizinischen oder personenbezogenen Kontext.

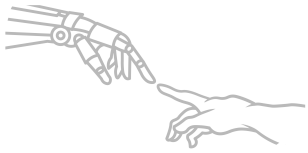
Spezielle Angriffsszenarien und ergänzende Schutzmaßnahmen Neben klassischen Bedrohungen wie adversarialen Beispielen oder Modell-Poisoning, sowie Daten-bezogene Risiken (vgl. 3.2) für überwachtes, halbüberachtes und unüberwachtes Lernen existieren Angriffsmuster, die gezielt Schwachstellen in speziellen Lernverfahren, Trainingsmethoden oder Evaluationsprozessen ausnutzen. Diese Angriffe entfalten ebenfalls potenziell erhebliche Wirkung auf Sicherheit, Verlässlichkeit und Datenschutz von KI-Systemen.

Reinforcement Learning

Zu solchen Lernverfahren-spezifischen Risiken im Bereich des Reinforcement Learning (RL) gehören beispielsweise Angriffsmuster, die sich auf die Wechselwirkung zwischen Agent und Umgebung konzentrieren. Beim *Reward Hacking* wird die Belohnungsfunktion so manipuliert, dass der Agent zwar scheinbar erfolgreich handelt, tatsächlich aber nicht das gewünschte Verhalten ausführt. Ein Agent, der Punkte für wiederholte Zustände erhält, könnte sich beispielsweise in Endlosschleifen optimieren. Beim *Exploration Hacking* wird das Explorationsverhalten gezielt gestört, sodass der Agent falsche Rückschlüsse zieht oder seine Strategieentwicklung frühzeitig abbricht. Schutz bieten hier robuste Reward-Definitionen, Simulation von Störungen in der Umgebung sowie ein kontinuierliches Monitoring des Lernfortschritts im Zeitverlauf.

Generative Modelle

Generative Modelle wie *GANs* weisen ebenfalls eigene Schwachstellen auf. Beim sogenannten *Mode Collapse* lernt das Modell nur noch sehr wenige Ausgabemuster zu erzeugen, wodurch Vielfalt und Aussagekraft stark eingeschränkt werden. In sicherheitsrelevanten Anwendungen kann dies dazu führen, dass generierte Inhalte wichtige Musterbereiche systematisch ignorieren. Verstärkt werden kann dieses Verhalten durch gezielte Eingriffe in die Trainingsrückkopplung, etwa durch manipulierte *Diskriminator-Ausgaben*. Eine gezielte Förderung der Vielfalt – etwa durch Regularisierung, Diversitätsmetriken oder angepasste Loss-Funktionen – trägt dazu bei, solchen Effekten vorzubeugen.



Federated Learning

Im Federated Learning entstehen neue Angriffsflächen durch die verteilte Trainingsstruktur. Beim *Byzantine Attack* sendet ein kompromittierter Client absichtlich fehlerhafte Modellupdates, um das globale Modell in eine falsche Richtung zu lenken. Noch wirkungsvoller ist der *Sybil Attack*, bei dem ein Angreifer mehrere virtuelle Clients einführt und damit die Aggregation gezielt dominiert. Gegenmaßnahmen bestehen in der Anwendung robuster Aggregationsverfahren wie *Krum* oder *Trimmed Mean* sowie in der Einführung von Authentifizierungsmechanismen zur Validierung der teilnehmenden Clients.

Hyperparameteroptimierung

Auch in der Phase der Hyperparameteroptimierung können gezielte Angriffe erfolgen. So können fehlerhafte Konfigurationen etwa durch manipulierte *AutoML*-Komponenten eingeschleust oder über sogenannte *Early-Stopping-Angriffe* das Training zu früh beendet werden. Weitere Bedrohungen entstehen durch *Randomized* oder *Brute-Force-Angriffe*, die den Suchraum gezielt ausweiten oder unbrauchbare Konfigurationen bevorzugen. Um die Integrität des Tuning-Prozesses zu sichern, sollten adaptive Optimierungsalgorithmen wie *Bayesian Optimization* mit Validierungsprotokollen kombiniert werden. Ergänzend bieten Logging, Zugriffsbeschränkungen und Tests zur Reproduzierbarkeit wirksame Schutzmechanismen.

Online-Lernen

Ein weiteres spezielles Szenario ergibt sich im Online-Lernen. Hier wird das Modell fortlaufend aktualisiert, was einerseits hohe Anpassungsfähigkeit ermöglicht, andererseits jedoch auch potenzielle Risiken birgt. Beim *Concept Drift Attack* verändern Angreifer die Datenverteilung schrittweise, sodass das Modell langsam in eine ungünstige Richtung verschoben wird. Da diese Form des Angriffs besonders subtil ist, sollte ein kontinuierliches Monitoring mit Drift-Erkennung eingesetzt werden, etwa durch Vergleich der aktuellen Vorhersagen mit Referenzmodellen oder durch Analyse der Veränderung von Gradienten über längere Zeiträume hinweg.

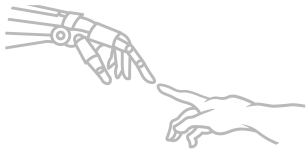
Die verschiedenen Lernverfahren, sowie mögliche spezifische Angriffsvektoren und Absicherungsmaßnahmen

sind zur Übersicht in Tabelle 3.2 aufgelistet.

Tabelle 3.2: Auswahl von Angriffsmustern mit spezifischen Schutzmaßnahmen entlang verschiedener Lernverfahren

Lernverfahren	Angriffsvektor	Absicherungsmaßnahmen
Reinforcement Learning	*Reward Hacking*: Manipulation der Belohnungsfunktion zur Erzeugung suboptimaler Strategien	▶ Robuste Reward-Definition ▶ Simulation unerwarteter Zustände ▶ Monitoring des Lernfortschritts
	Exploration Hacking: Störung des Erkundungsverhaltens zur Verhinderung sinnvoller Strategien	▶ Analyse von Erkundungsmustern ▶ Evaluation unter alternativen Umgebungen ▶ Kontrollierte Erkundungseinschränkung
Generative Modelle (GANs)	*Mode Collapse*: Verengung der Ausgabemuster durch überoptimierte Diskriminatorrückmeldung	▶ Diversitätsmetriken ▶ Entropie-basierte Regularisierung ▶ Anpassung der Verlustfunktion
Federated Learning	*Byzantine Attack*: absichtlich fehlerhafte Modellupdates zur Sabotage des globalen Modells	▶ Robuste Aggregationsverfahren (z.B. Krum, Trimmed Mean) ▶ Anomalieprüfung von Client-Updates
	Sybil Attack: Nutzung mehrerer gefälschter Clients zur Manipulation der Aggregation	▶ Client-Authentifizierung ▶ Reputationsbewertung ▶ Identitätsprüfung durch Zertifikate
Hyperparameteroptimierung AutoML	*Early-Stopping-Angriff*: absichtlicher vorzeitiger Abbruch des Trainingsprozesses	▶ Validierung mit Kontrollmetriken ▶ Überwachung des Abbruchverhaltens ▶ Logging von Trainingsverläufen
	Randomized Attack: Störung des Suchraums durch gezielte Unschärfe	▶ Adaptive Optimierung (z.B. Bayesian Optimization) ▶ Eingrenzung des Parameterraums ▶ Plausibilitätsprüfung von Ergebnissen
	Brute-Force-Angriff: Überlastung oder gezielte Desorientierung der Tuninglogik	▶ Ressourcenschutz durch Limitierung der Versuche ▶ Monitoring des Optimierungsverlaufs ▶ Reproduzierbarkeitstests
Online-Lernen	*Concept Drift Attack*: schleichende Veränderung der Datenverteilung zur Modellverzerrung	▶ Kontinuierliche Drift-Erkennung ▶ Vergleich mit Referenzmodellen ▶ Kontrolle über Zeitreihenverhalten

Modellzugriff und Schutz vor Missbrauch Nach dem Training und der Integration eines Modells in ein KI-System stellt sich eine zentrale sicherheitsrelevante Frage: Wer darf das Modell wann, wie und in welchem Umfang nutzen? Unzureichend gesicherte Modellzugriffe können nicht nur zu Datenabflüssen, sondern auch zu gezielten Manipulationen, Ausforschung oder unbemerkter Rekonstruktion vertraulicher Informationen



führen. Daher ist es unerlässlich, frühzeitig technische und organisatorische Maßnahmen zu etablieren, die eine kontrollierte und abgesicherte Nutzung des Modells sicherstellen.

Zugriffskontrolle

Ein grundlegender Baustein ist auch im Kontext des Modells die Zugriffskontrolle. Modelle – insbesondere solche, die personenbezogene Daten oder geschäftskritisches Wissen verarbeiten – sollten niemals öffentlich zugänglich oder ungeschützt ausgeliefert werden. Es ist daher ratsam, die in Abschnitt Datenerfassung erwähnten Mechanismen zur Zugriffskontrolle auch für das KI-Modell umzusetzen.

Rekonstruktion des Modells

Doch nicht nur wer zugreift, sondern auch was genau abgefragt wird, ist entscheidend. Über sogenannte *Model Extraction Attacks* können Angreifer durch systematische Abfragen ein Modell schrittweise rekonstruieren – insbesondere, wenn dieses detaillierte Wahrscheinlichkeitswerte oder Gradienteninformationen zurückgibt. Deshalb sollte die Ausgabe des Modells in sicherheitssensitiven Kontexten bewusst begrenzt werden, etwa durch Rückgabe nur der Top-K-Klassen oder durch gerundete Scores. Auch *Rate Limiting* für API-Zugriffe sowie Monitoring-Mechanismen zur Erkennung ungewöhnlicher Zugriffsmuster tragen zum Schutz bei.

Modelldiebstahl

Neben dem schrittweisen Rekonstruieren von Modellen über API-Zugriffe – wie bei *Model Extraction Attacks* – besteht ein weiteres, unmittelbares Risiko: der direkte Diebstahl eines trainierten Modells. Dabei verschaffen sich Angreifer Zugriff auf gespeicherte Modellartefakte, beispielsweise in Form von .pt- oder .onnx-Dateien, oder extrahieren Modellparameter direkt aus dem Hauptspeicher eines laufenden Systems. Anders als bei rekonstruktiven Angriffen erfolgt hier ein direkter Bruch in das System oder die Entwicklungsumgebung, um das Modell vollständig zu kopieren.

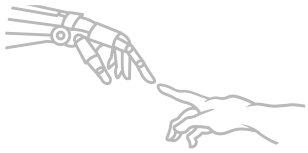
Ein typischer Angriffsvektor ist der ungeschützte Zugriff auf Dateisysteme in Cloud-Umgebungen, auf denen Modelle abgelegt sind. Auch veraltete Backup-Systeme oder

falsch konfigurierte Objekt-Speicher (z.B. offen zugängliche *S3-Buckets*) bieten oft unbeabsichtigt Zugang zu sensiblen Modellartefakten. Selbst temporäre Artefakte, die bei der Konvertierung oder Quantisierung erstellt werden, können ausreichend Informationen enthalten, um ein Modell vollständig zu reproduzieren oder in einen fremden Kontext zu übertragen.

In anderen Fällen nutzen Angreifer *Side-Channel Techniken*, um Modellparameter direkt aus dem Arbeitsspeicher auszulesen. Besonders bei Deployment in geteilten Infrastrukturen oder nicht abgesicherten Edge-Geräten (wie in Smart-Home-Systemen oder Fahrzeugen) besteht hier ein relevantes Risiko.

Zur Absicherung gegen direkten Modelldiebstahl sind die in Abbildung 3.5 aufgezählten Strategien zu empfehlen.

Abbildung 3.5: Abwehrstrategien gegen Modelldiebstahl.



Ableitung sensibler Informationen

Ein weiterer Aspekt ist der Schutz vor sogenannten *Membership Inference Attacks*. Hierbei versuchen Angreifer herauszufinden, ob ein bestimmter Datensatz Teil des Trainings war – was in vielen Fällen datenschutzrechtlich hochrelevant ist. Abhilfe schaffen auch hier Techniken wie *Differential Privacy*, die gezielt Rauschen in das Modellverhalten integrieren, ohne dessen Gesamtleistung wesentlich zu beeinträchtigen.

Kontrollierte Umgebungen

Ergänzend ist es ratsam, kritische Modelle nur in kontrollierten Umgebungen auszuführen – z.B. innerhalb isolierter Container mit restriktiven Netzwerkrichtlinien oder in Hardware-gestützten sicheren Ausführungsumgebungen. So wird verhindert, dass das Modell oder seine Ausgaben während der Laufzeit abgefangen oder manipuliert werden können.

3.4 Produktivsetzung

Die Produktivsetzung von KI-Systemen ist eine kritische Phase, bei welcher weitere sicherheitsrelevante Maßnahmen ergriffen werden sollten. Dieser Schritt stellt besondere Anforderungen an die Cybersicherheit, da nun nicht mehr nur Trainingsdaten oder Modelle, sondern komplette Systemkomponenten potenziell angreifbar werden. Anders als bei klassischen Softwareprodukten ergibt sich die Angriffsfläche bei KI-Systemen oft aus einem Zusammenspiel aus Modellen, Schnittstellen, Infrastruktur, Automatisierungspipelines und operativen Prozessen. Ziel der Absicherung ist es daher, eine kontrollierte, überprüfbare und widerstandsfähige Überführung in den produktiven Betrieb zu ermöglichen.

Vorbereitung der Infrastruktur

Ein erster Schwerpunkt liegt auf der Vorbereitung der zugrundeliegenden Infrastruktur. Hierzu zählt die Absicherung der technischen Umgebung, in der das KI-System künftig ausgeführt wird – sei es in Form von Containern, virtuellen Maschinen oder verwalteten Cloud-Diensten. Systeme sollten gehärtet, minimal konfiguriert und auf unnötige Dienste oder offene Ports geprüft werden. Hierzu gehört ebenfalls die Trennung von

Trainings-, Test- und Produktionsumgebung. Eine Bereitstellung mittels Konzepten wie *Infrastructure-as-Code* erlaubt es, die Konfiguration reproduzierbar und versioniert zu dokumentieren. Auch müssen CI/CD-Pipelines gegen unautorisierte Änderungen geschützt werden. Dies umfasst den Schutz von Build-Skripten, Zugangskontrollen zu Repositories und die Sicherstellung, dass nur geprüfte und signierte Artefakte in produktionsnahe Umgebungen übernommen werden. Somit unterliegt das Deployment selbst einem Genehmigungsprozess und sollte ausschließlich durch autorisierte Rollen ausführbar sein.

Zugriffskonzept

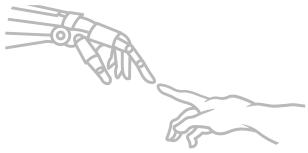
Parallel dazu sollte das Zugriffskonzept für das Gesamtsystem vor der Produktivsetzung überprüft und angepasst werden. Es ist sicherzustellen, dass alle internen und externen Systemzugriffe über zentrale Authentifizierungsmechanismen laufen und Rechte nach dem Prinzip der minimalen Berechtigung vergeben sind. Insbesondere Deployment-, API- und Infrastrukturzugriffe sollten durch rollenbasierte Modelle getrennt sein. Eine Multi-Faktor-Authentifizierung ist für alle privilegierten Rollen obligatorisch. Um ungewollte Änderungen nachvollziehen zu können, ist die Protokollierung aller administrativen und sicherheitsrelevanten Aktionen bereits in dieser Phase sinnvoll.

Externe Schnittstellen

Ein besonderes Augenmerk gilt der Absicherung externer Schnittstellen. Vor allem KI-gestützte APIs, die Inferenzdienste oder Feedbackannahmen bereitstellen, bilden ein häufig genutztes Einfallstor. Noch vor der Freischaltung müssen sie auf Authentifizierung, Eingabeverhalten und Rate-Limiting geprüft werden. Eingehende Daten sollten strukturell validiert und gegen typische Angriffsformen wie Injection, unzulässige Serialisierung oder Typmanipulation geschützt sein. Der Einsatz eines vorgeschalteten API-Gateways mit Transportverschlüsselung, Zugriffskontrolle und Logging-Funktionalität kann helfen, Angriffe frühzeitig zu erkennen und abzuwehren.

Systembestandteile

Neben Code- und API-Ebene ist die Herkunft und Integrität sämtlicher Systembestandteile sicherzustellen.



Dies betrifft nicht nur das Modell selbst, sondern auch Konfigurationsdateien, Umgebungsparameter, Third-Party-Bibliotheken und eingesetzte Container. Alle diese Komponenten sollten vor der Produktivsetzung signiert oder mit Hashwerten versehen und in einem Release-Protokoll dokumentiert werden. Wie in Abschnitt 3.2 beschrieben, müssen Abhängigkeiten aus externen Quellen auf Vertrauenswürdigkeit geprüft, Pakete regelmäßig aktualisiert und Supply-Chain-Risiken systematisch bewertet werden. Dynamische Nachladungen von Bibliotheken oder Modellen aus nicht überwachten Quellen sollten grundsätzlich ausgeschlossen werden.

Modell-Ausgaben

Auch die Rückgabeformate des Modells bergen sicherheitsrelevante Fragestellungen. Liefert ein Modell etwa vollständige Wahrscheinlichkeitsverteilungen oder Confidence Scores, steigt die Angriffsfläche für Inference-basierte Attacken. In sicherheitskritischen Szenarien kann es daher sinnvoll sein, die Modellantwort auf Top-K-Ergebnisse zu beschränken oder Score-Werte ganz zu unterdrücken. In Szenarien mit hohem Schutzbedarf – etwa im medizinischen oder regulatorischen Bereich – sollten zusätzlich strukturierte Reviews der Vorhersageausgaben („Human-in-the-Loop“) erfolgen, um unbeabsichtigte Informationslecks zu vermeiden.

Security Assessment des Gesamtsystems

Im Vorfeld der Freigabe empfiehlt sich ein gezieltes Security-Assessment des gesamten Systems. Dieses kann technische Prüfungen wie Penetrationstests oder statische Analysen ebenso umfassen wie ein strukturiertes Threat Modeling. Dabei werden potenzielle Angriffsvektoren auf Systemebene identifiziert und bewertet – etwa im Hinblick auf API-Exponierung, Identitätsverwaltung oder seitliche Bewegungen im Netzwerk. Auf dieser Basis können Eingangskriterien für die Freigabe definiert werden, zum Beispiel vollständige Auditprotokolle, nachgewiesene Ausfallsicherheit oder dokumentierte Kommunikationspfade zwischen Modulen. Auch ein technisches Rollback-Szenario sollte vorab vorbereitet sein, um im Fehlerfall schnell auf eine letzte stabile Version zurückkehren zu können.

Noch vor der eigentlichen Produktivsetzung sollten KI-

Systeme systematisch auf technische Schwächen und Risiken geprüft werden. Hierzu empfiehlt sich die Integration automatisierter Testverfahren in bestehende CI/CD-Pipelines. Der Vorteil dieser Einbettung liegt in der Reproduzierbarkeit und frühen Erkennung potenzieller Probleme bereits im Build- oder Integrationsprozess. So lassen sich sicherheitsrelevante Auffälligkeiten, Performanceeinbrüche oder unerwartetes Modellverhalten erkennen, bevor ein Release in produktionsnahe Umgebungen gelangt. Die Durchführung dieser Tests kann automatisiert erfolgen, die Auswertung sollte jedoch durch entsprechend geschulte Test-, Audit- oder KI-Sicherheitsteams begleitet werden.

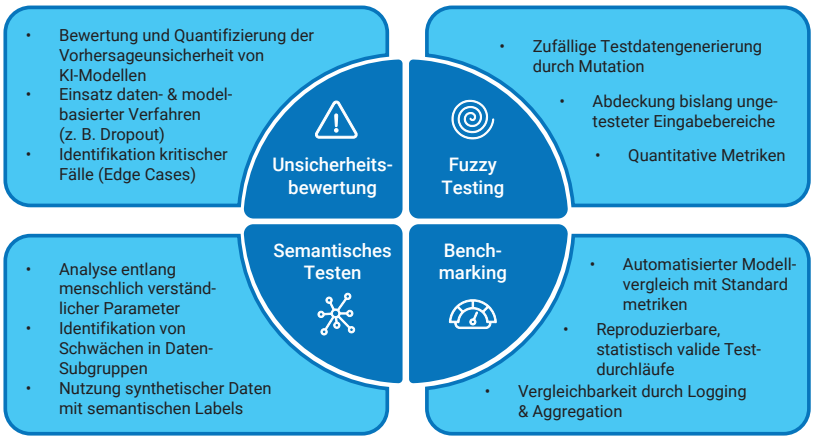


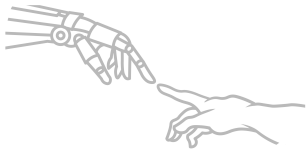
Abbildung 3.6: Verfahren zur systematischen Bewertung der Sicherheit im Kontext von Vertrauensaspekten von KI-Systemen.

Benchmarking

In der Praxis haben sich vier Kategorien von Prüfmethoden bewährt, die unterschiedliche Aspekte der Vertrauenswürdigkeit und Sicherheit abdecken (siehe Abbildung 3.6). Ein Verfahren ist das Benchmarking, bei dem verschiedene Modellversionen anhand definierter Metriken verglichen werden. Durch reproduzierbare Tests mit Logging und Aggregation kann beispielsweise sichergestellt werden, dass eine neue Modellversion nicht unbeabsichtigt an Genauigkeit verliert oder sich in bestimmten Szenarien regressiv verhält. Solche Performance-Vergleiche bilden eine objektive Grundlage für Freigabeentscheidungen.

Unsicherheitsbewertung

Ergänzend dazu wird zunehmend auf Unsicherheitsbewertung gesetzt. Diese zielt darauf ab, wie „selbstsicher“ das Modell in seinen Vorhersagen ist und ob es in kritischen Situationen einen potenziellen Fehler kommuni-



zieren kann. Dabei kommen daten- und modellbasierte Methoden zum Einsatz – etwa *Bayesian Dropout* oder *Ensemble-Ansätze*. In sicherheitskritischen Kontexten wie autonomem Fahren oder medizinischer Diagnose ist das gezielte Identifizieren und Markieren sogenannter *Edge Cases* für eine spätere menschliche Überprüfung notwendig. Modelle, die Unsicherheiten transparent kommunizieren, ermöglichen eine informierte Risikoabwägung im Betrieb.

Fuzzy Testing

Neben Performance und Unsicherheit spielt die Robustheit des Modells gegenüber unerwarteten oder manipulierten Eingaben eine entscheidende Rolle. Zwei besonders praxisnahe Methoden stehen hierfür zur Verfügung: Fuzzy Testing und Semantisches Testen. Beim Fuzzy Testing werden Eingaben systematisch durch Zufallsmutationen verändert. Ziel ist es, ungewöhnliche oder selten getestete Modellzustände zu erreichen und dabei fehleranfälliges Verhalten offenzulegen. Diese Methode stammt ursprünglich aus dem Bereich der Software-Sicherheit und wurde erfolgreich auf KI-Systeme übertragen. Besonders geeignet ist Fuzzy Testing für Anwendungsfälle mit großen, unstrukturierten Eingangsräumen – etwa in der Bild- oder Textverarbeitung. Über quantitative Metriken wie die sogenannte *Neuron Coverage* lässt sich die Testtiefe messbar machen.

Semantisches Testen

Das Semantische Testen geht einen Schritt weiter, indem es gezielt entlang menschlich verständlicher Parameter prüft. Statt rein zufälliger Variationen analysiert dieses Verfahren, wie das Modell auf inhaltlich relevante Unterschiede reagiert – beispielsweise Fahrzeugfarbe bei Objekterkennung oder Dialektvariationen in Sprachmodellen. Durch die Nutzung synthetisch erzeugter Testdaten mit expliziten Attributen können Schwächen in bestimmten Subgruppen gezielt identifiziert werden. Solche systematischen Tests helfen, Risiken wie algorithmische Benachteiligung oder *Hidden Biases* frühzeitig aufzudecken.

Betriebsmonitoring

Die Übergangsphase in den Betrieb sollte zudem von einem gezielten Betriebsmonitoring begleitet werden. Noch bevor das System offiziell in den Betrieb überführt

wird, sollten grundlegende Telemetriekomponenten eingerichtet sein – etwa zur Überwachung von Antwortzeiten, Ressourcennutzung, Systemstatus und Zugriffsmustern. Sicherheitsrelevante Metriken wie Authentifizierungsfehler, API-Auslastung oder nicht autorisierte Konfigurationsversuche sollten bereits in der Produktivsetzungsphase zentral erfasst und ausgewertet werden. Die Einbindung in bestehende SIEM-Systeme (Security Information and Event Management) kann die Alarmierung und forensische Analyse im Fall eines Vorfalls erleichtern.

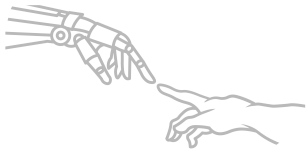
Organisatorisches Umfeld

Auch das organisatorische Umfeld muss auf den Produktivbetrieb vorbereitet sein. Rollen und Zuständigkeiten – etwa für Incident Response, Betrieb, Freigabe und Monitoring – sollten klar definiert sein. Beteiligte Teams sollten über spezifische Risiken KI-basierter Systeme informiert und in der Lage sein (vgl. Kompetenzen [12]), technische und sicherheitsrelevante Besonderheiten zu erkennen. Eine begleitete Übergabe mit 24- oder 48-stündiger Monitoring-Verantwortung kann dabei helfen, unerwartetes Verhalten frühzeitig zu identifizieren und Maßnahmen einzuleiten.

3.5 Betrieb

Nach der Produktivsetzung beginnt die Betriebsphase – und mit ihr die kontinuierliche Verantwortung für die Cybersicherheit des Gesamtsystems. Anders als in der Entwicklung, wo Tests meist in kontrollierten Umgebungen erfolgen, agiert das KI-System nun unter realen Bedingungen mit echten Nutzern, dynamischen Datenströmen und potenziellen Angreifern. Sicherheit ist hier kein Zustand, sondern ein Prozess: Der Betrieb muss auf Veränderungen reagieren, Schwächen erkennen und neue Bedrohungen abwehren – automatisiert, nachvollziehbar und belastbar.

Angriffsvektoren Neben klassischen Angriffsvektoren auf IT-Systeme, wie etwa einem Denial-of-Service-



Angriff (DoS) sind KI-Systeme während ihres Betriebs einer Vielzahl spezifischer Angriffsszenarien ausgesetzt, die über klassische IT-Bedrohungen hinausgehen. Einige dieser Angriffsvektoren betreffen primär die Daten oder das Modell und wurden in den Abschnitten 3.2 und 3.3 beschrieben (z.B. Daten-Poisioning). Diese können in der Operationalisierung in einem anderen Kontext aufkommen. Spezifisch für die Betriebsphase sind folgende prominente Angriffsvektoren bekannt.

Eingabemanipulationen (Input Poisoning) Angreifer platzieren gezielt Daten mit schädlicher Struktur, um das Modell zu täuschen oder Fehlklassifikationen hervorzurufen – etwa durch manipulierte Sensorwerte oder gefälschte Texteingaben.

Evasion-Angriffe Hierbei werden Eingaben subtil verändert (z.B. durch Bildrauschen oder syntaktische Variationen), sodass das Modell eine falsche Vorhersage trifft, obwohl der Mensch die Eingabe weiterhin korrekt klassifizieren würde.

Prompt Injection Besonders bei generativen KI-Systemen (z.B. LLMs) schleusen Angreifer spezielle Sprachmuster ein, um unerlaubte Antworten oder Sicherheitslecks zu provozieren.

Model Extraction Angreifer rekonstruieren das Modell durch viele API-Anfragen – etwa zur Replikation des Verhaltens oder zur Analyse von Schwachstellen.

Schutzmaßnahmen Diesen vielfältigen Risiken begegnet man am besten durch eine Kombination aus technischen Schutzmaßnahmen, organisatorischen Prozessen und kontinuierlichem Lagebewusstsein.

Eingabevalidierung Eine wichtige erste Verteidigungslinie bildet die Eingabevalidierung: Alle Daten, die in das System gelangen, müssen strukturell und semantisch geprüft werden – sei es auf Typ, Format, Wertebereiche oder auffällige Muster. Das gilt insbesondere für Schnittstellen wie APIs, Sensor-Gateways oder Nutzereingaben. Ergänzend sollte ein API-Gateway vorgeschaltet sein, das Transportverschlüsselung, Rate Limiting und Logging sicherstellt.

Ausgabevalidierung

Im Fall generativer Systeme empfiehlt sich eine Ausgabevalidierung: Bei Modellen, die Sprache oder strukturierte Daten erzeugen, sollte geprüft werden, ob darin ungewollte Inhalte, sensible Informationen oder steuerbare Tokens enthalten sind. So kann etwa ein LLM-Ausgang durch Blacklisting oder ebenfalls Top-K-Filterung beschränkt werden, um das Risiko von Prompt Injection und Halluzinationen zu reduzieren.

Zugangskontrolle

Ein weiterer Baustein ist die Zugangskontrolle auf System-, API- und Modell-Ebene. Zugriffe auf inferenzfähige Modelle sollten stets authentifiziert und autorisiert sein – idealerweise durch RBAC, MFA und tokenbasierte Authentifizierung. Debug- und DevOps-Schnittstellen müssen gesondert geschützt oder abgeschaltet werden. Auch Monitoring-Tools und Dashboards mit Modellinformationen sollten nicht öffentlich verfügbar sein.

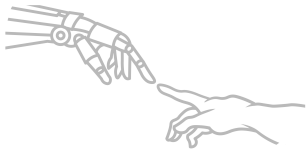
Monitoring

Ein weiteres Element ist das laufende Monitoring von betriebs- und sicherheitsrelevanter Metriken. Neben klassischen Betriebskennzahlen wie Antwortzeit, CPU-Auslastung oder Netzwerklast müssen auch KI-spezifische Größen erfasst werden – etwa Confidence Scores, Distribution der Eingabedaten oder Frequenzen einzelner Modellklassen. Abweichungen von erwarteten Mustern können auf Daten- oder Modell-Drifts hinweisen, aber auch auf gezielte Angriffsversuche.

Für robuste Betriebsüberwachung eignen sich neben klassischem Telemetrie-Monitoring auch Intrusion Detection Systeme (IDS) oder verhaltensbasierte Anomalieerkennungen. Letztere können etwa ungewöhnliche Modellantworten, stark veränderte Zugriffsmuster oder fehlerhafte API-Nutzungen erkennen – und automatisierte Reaktionen auslösen. Wichtig ist dabei ein klar definierter Eskalationspfad, der von Alarmierung über temporäre Deaktivierung bis hin zum technischen Rollback reicht.

Regelmäßiges Nachtrainieren

Ein weiterer Aspekt im sicheren Betrieb ist der Umgang mit Modell-Updates. Regelmäßiges Nachtrainieren ist eine Möglichkeit, um Drifts entgegenzuwirken – muss jedoch unter streng kontrollierten Bedingungen erfolgen.



Jede neue Modellversion sollte in Re-Deploymentphase automatisch getestet, beispielsweise die in Abbildung 3.6 dargestellten Tests, dokumentiert und durch Autorisierung freigegeben werden. Feedback-Schleifen aus Nutzereingaben können hierbei helfen, das Modell zu verbessern, dürfen aber nie unkontrolliert in das Trainingsset einfließen – Stichwort: Poisoning durch Feedback Loops.

3.6 Abschaltung

Auch wenn ein KI-System nicht mehr produktiv genutzt wird, bleiben sicherheitsrelevante Risiken bestehen, die bei der Abschaltung gezielt adressiert werden müssen. Oft unterschätzt, kann diese Phase zur kritischen Schwachstelle im Lebenszyklus werden – insbesondere dann, wenn Daten, Modelle oder Systemreste unkontrolliert zurückbleiben. Ziel der Abschaltung ist es daher, das System dauerhaft zu deaktivieren, alle sicherheitsrelevanten Artefakte zu entfernen und die potenzielle Wiederverwendbarkeit unter Kontrolle zu bringen. Um einen geregelten und sicheren Abschaltprozess zu gewährleisten, sind verschiedene Schritte notwendig, die zur Übersicht in Abbildung 3.7 dargestellt sind und nachfolgend erläutert werden.

Vermeidung von Datenlecks

Ein mögliches Risiko in dieser Phase sind Datenlecks durch unzureichend gelöschte Systemkomponenten. Gerade KI-Modelle enthalten oft Trainingsdaten, eingebettete Parameter oder abgeleitete Informationen, die Rückschlüsse auf sensible Inhalte erlauben. Wo keine

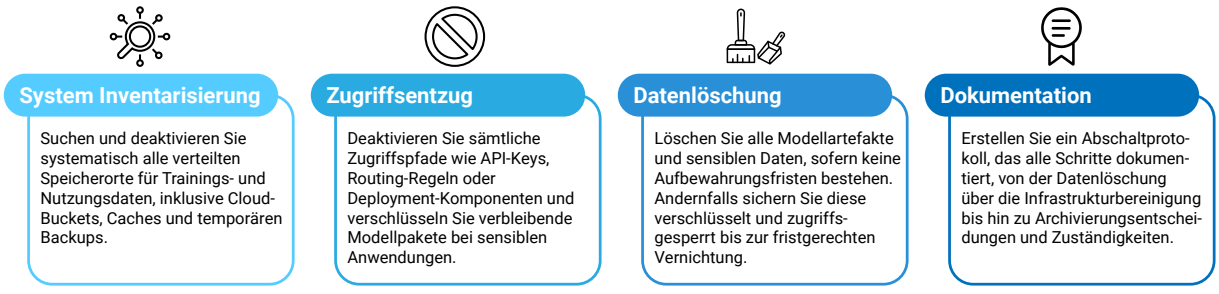
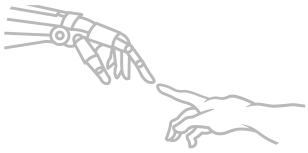


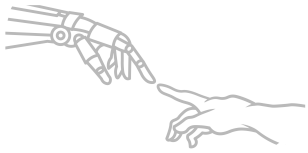
Abbildung 3.7: Übersicht über zentrale Aspekte bei der sicheren Abschaltung von KI-Systemen.



rechtlichen Aufbewahrungspflichten bestehen, sollten diese Informationen sicher gelöscht werden – etwa durch zertifiziertes Überschreiben oder physische Vernichtung der Speichermedien. Was das KI-Modell betrifft ist es wichtig, alle Modellartefakte – insbesondere gespeicherte Checkpoints, Serialisierungen (z.B. `.pt`, `.pb`, `.h5`) und begleitende Konfigurationsdateien – vollständig und nachvollziehbar zu entfernen. Je nach Sensibilität der Daten empfiehlt sich nicht nur die logische Löschung, sondern auch das mehrfache Überschreiben (z.B. nach DoD-Standard) oder physische Vernichtung der betroffenen Datenträger. Falls gesetzliche oder betriebliche Aufbewahrungsfristen zu beachten sind, müssen die betroffenen Daten bis zur endgültigen Löschung technisch gesichert, verschlüsselt und für den Zugriff gesperrt archiviert werden.

Darüber hinaus muss geprüft werden, wo und in welcher Form Trainings- oder Nutzungsdaten im System gespeichert wurden – etwa in Datenbanken, Zwischenspeichern, Logging-Systemen oder Backup-Dateien. Besonders bei verteilten Systemen ist es ratsam, systematisch nach potenziellen Speicherorten zu suchen, die außerhalb des eigentlichen Hauptsystems liegen – etwa in Cloud-Buckets, CDN-Caches oder Remote-Volumes. Die Einführung einer *Ablage-Policy* für das Modelltraining kann schon im Betrieb helfen, spätere Löschprozesse zu vereinfachen.

Besondere Sorgfalt ist auch bei der Wiederverwendung von Hardware und Software erforderlich. Systeme, die im produktiven KI-Betrieb standen – etwa Trainingsrechner, Edge-Geräte oder Containerumgebungen – können nach wie vor sensible Informationen enthalten. Vor einer Weiternutzung sollten Speichermedien vollständig gelöscht oder zertifiziert überschrieben werden. Falls die Infrastruktur von Drittanbietern bereitgestellt wurde (z.B. Cloud-Instanzen), empfiehlt sich die schriftliche Bestätigung der sicheren Entsorgung durch den Anbieter. Auch temporär verwendete Container-Images, virtuelle Maschinen oder Snapshots sollten vollständig aus Repositories und Speicherorten entfernt werden.

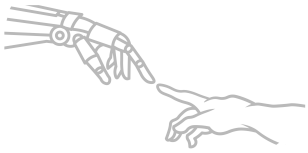


Vermeidung unbefugter Zugriffe

Ein weiterer Aspekt stellt der unbefugte Zugriff auf ein deaktiviertes, aber technisch noch vorhandenes KI-Modell dar. In der Praxis kann es vorkommen, dass abgeschaltete Modelle weiterhin über APIs erreichbar sind oder durch administrative Fehler versehentlich im System verbleiben. Es ist daher wichtig, nicht nur die Anwendung zu beenden, sondern auch sämtliche Zugriffsmöglichkeiten technisch auszuschließen. Dazu zählen: Deaktivierung von API-Keys, Abschaltung von Routing-Regeln und vollständige Entfernung der Modell-Deployment-Komponenten aus produktionsnahen Pipelines. Für sensible Anwendungen ist zusätzlich die Verschlüsselung oder Obfuskation verbleibender Modellpakete ratsam, um Missbrauch durch Reengineering zu erschweren.

Dokumentation

Zur Sicherstellung der Nachvollziehbarkeit empfiehlt sich die Erstellung eines standardisierten Abschaltprotokolls. Darin sollten alle Schritte festgehalten werden – etwa deaktivierte Services, gelöschte Daten, entfallene Rollen, genutzte Werkzeuge zur Datenvernichtung sowie verbleibende Archivierungsentscheidungen. Dieses Dokument erfüllt nicht nur interne Audit- und Compliance-Anforderungen, sondern dient auch als Blaupause für zukünftige Stilllegungen. In regulierten Bereichen – etwa im medizinischen, öffentlichen oder kritischen Infrastrukturbereich – ist eine formelle Genehmigung der Abschaltung durch autorisierte Rollen empfehlenswert.

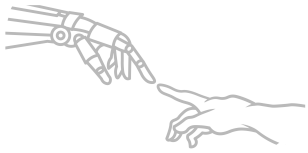


Governance 4

Mit dem zunehmenden Einsatz künstlicher Intelligenz in unternehmenskritischen Prozessen entsteht die Notwendigkeit, bestehende Cybersicherheitsstrukturen gezielt weiterzuentwickeln. Klassische Schutzkonzepte reichen nicht aus, um den besonderen Anforderungen intelligenter Systeme gerecht zu werden. Daher muss die Security-Governance eines Unternehmens um spezifische Prinzipien, Mechanismen und Rollen erweitert werden, die den Umgang mit KI-Risiken systematisch steuern.

In Kapitel 3 haben wir die technologischen Aspekte sowie deren Umsetzung entlang des Produktlebenszyklus von KI-Systeme betrachtet. Es hat sich gezeigt, dass ergänzend zu etablierten IT-Schutzzielen auch KI-spezifische Implementierungen in Bezug auf die Cybersicherheit bedacht werden müssen. Dieser Wandel betrifft jedoch nicht nur die KI-Systeme auf einer technischen Ebene, sondern erfordert auch organisatorische Anpassungen. Beispielsweise sollten die Zuständigkeiten für den sicheren Betrieb von KI-Systemen klar geregelt sein – insbesondere, wenn mehrere Teams wie Data Science, MLOps und Security beteiligt sind. Um Reibungsverluste zu vermeiden, empfiehlt es sich, ebenfalls dedizierte Rollen für die KI-Sicherheit zu schaffen, etwa in Form spezialisierter KI-Security-Officer oder fachübergreifender Taskforces. Die Einbindung dieser Funktionen in bestehende ISMS-Strukturen schafft Klarheit und fördert auch die ganzheitliche Steuerbarkeit des Themas.

Damit diese Strukturen wirksam greifen, sind verbindliche Leitplanken notwendig. Anstelle punktueller Ad-hoc-Maßnahmen braucht es konsistente Richtlinien, die sicherstellen, dass Sicherheit in allen Phasen der KI-Entwicklung und -Nutzung mitgedacht wird. Hierzu gehören klare Regeln zur Data Governance – etwa zur Herkunft, Qualität und Absicherung von Trainingsdaten – ebenso wie Vorgaben für sicheres Modelltraining,



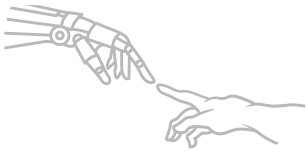
Modellhärtung und Versionierung. Als Orientierung können etablierte Standards wie ISO/IEC 27001, das NIST AI Risk Management Framework oder die Vorgaben der EU-KI-Verordnung dienen. Entscheidend ist, dass sie nicht nur formal berücksichtigt, sondern auch praktisch operationalisiert werden.

Eine effektive Governance setzt zudem voraus, dass die Bedrohungslage kontinuierlich neu bewertet wird. Für KI bedeutet das: klassische Risikoanalysen reichen nicht aus. Vielmehr müssen Schwachstellen systematisch aus Sicht von KI-Systemen identifiziert werden. Dabei ist nicht nur das KI-Modell selbst, sondern das gesamte Ökosystem – von der Datenquelle bis zur API – Gegenstand der Bewertung.

Um die technische Implementierung der Cybersicherheit in das Unternehmen einzubetten sind außerdem entsprechende Kompetenzen notwendig. Wenn KI-Sicherheit nachhaltig verankert werden soll, müssen alle beteiligten Akteure ein Grundverständnis für potenzielle Risiken und deren Abwehr entwickeln. Ebenso ist es sinnvoll, auch bestehende Blue- und Red-Teams auf KI-spezifische Angriffsszenarien vorzubereiten, um systematisch Sicherheitslücken aufzudecken – etwa durch simulierte Angriffe auf Testsysteme.

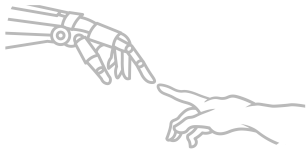
In sicherheitskritischen Anwendungsbereichen – etwa im medizinischen Umfeld oder in der öffentlichen Verwaltung – verschärfen sich die Anforderungen zusätzlich durch ethische Fragestellungen. Hier geht es nicht nur um Schutz vor Angriffen, sondern auch um die Vermeidung struktureller Diskriminierung, den Schutz sensibler personenbezogener Informationen oder die Sicherstellung von Transparenz in automatisierten Entscheidungen. Governance-Strukturen müssen diesen Anforderungen gerecht werden – sei es durch Ethik-Boards, Prüfverfahren oder technische Vorkehrungen wie differenzielle Privatsphäre.

Der langfristige Schutz von KI-Systemen fordert schließlich eine kontinuierliche Überprüfung und Anpassung. Dazu zählen nicht nur regelmäßige Sicherheitsaudits,



sondern auch ein ganzheitliches Monitoring während des Betriebs. Sicherheitsvorfälle sollten nicht nur erkannt, sondern auch durch spezifische Reaktionspläne abgefangen werden können – einschließlich der Möglichkeit, Modelle gezielt zurückzusetzen, zu isolieren oder reproduzierbar neu zu trainieren. Dieser Aspekt gewinnt umso mehr an Bedeutung, als sich sowohl Technologien als auch Bedrohungslagen dynamisch weiterentwickeln.

Vor diesem Hintergrund darf die Einbettung von KI-Sicherheit in die unternehmensweite Governance nicht als einmalige Maßnahme verstanden werden. Vielmehr handelt es sich um einen kontinuierlichen Prozess, der Lernbereitschaft, klare Zuständigkeiten und strategische Weitsicht erfordert. Wer frühzeitig entsprechende Strukturen etabliert, schafft nicht nur Sicherheit, sondern stärkt auch das Vertrauen in den nachhaltigen Einsatz von KI im Unternehmen.



5 Sicherheitskompetenz

In unserem Leitfaden „KI-Kompetenzen – Ein praktischer Leitfaden im Sinne der KI-Verordnung der EU“ [12] definieren wir fünf zentrale Kompetenzarten für den verantwortungsvollen Umgang mit KI-Systemen: Fachkompetenz, juristische Kompetenz, ethisch-reflexive Kompetenz, Datenkompetenz und technische KI-Kompetenz.

Innerhalb dieser Struktur finden sich auch erste sicherheitsrelevante Aspekte – etwa der Schutz personenbezogener Daten im Rahmen der Datenkompetenz oder die Berücksichtigung von Systemverhalten und Robustheit in der KI-technischen Kompetenz. Diese Inhalte zielen jedoch vorrangig auf ein grundlegendes Verständnis im jeweiligen Anwendungskontext beispielsweise zur sicheren Nutzung oder zur Vermeidung von Fehlfunktionen.

Ein durchgängiger und systematischer Umgang mit Sicherheitsanforderungen – wie ihn Artikel 15 der KI-Verordnung [1] fordert – geht jedoch über solche spezifischen sicherheitsbezogenen Kenntnisse und Fähigkeiten in anderen Kompetenzarten hinaus. Nur wenn Sicherheitskompetenz umfassend und systematisch aufgebaut und angewendet wird, können die rechtlichen und normativen Vorgaben erfüllt werden, die eine kontinuierliche Berücksichtigung von Robustheit, Genauigkeit und Cybersicherheit über den gesamten Lebenszyklus von KI-Systemen hinweg verlangen.

[1]: Das Europäische Parlament und der Rat der Europäischen Union (2024), *EU Artificial Intelligence Act*

Dies bestätigen auch die in Kapitel 2.2 beschriebenen etablierten Normen und Empfehlungen, die Cybersicherheit als kontinuierliche Querschnittsaufgabe begreifen – von der Entwicklung über Betrieb und Wartung bis zur Abschaltung eines KI-Systems. Daraus ergibt sich die Notwendigkeit einer übergreifenden, strukturierten Sicherheitskompetenz, die über ein Teilverständnis in spezifischen Kompetenzarten hinausgeht.

Sicherheitskompetenz umfasst Kenntnisse und Fähigkeiten, um sicherheitsrelevante Anforderungen im Umgang mit KI-Systemen zu verstehen, zu bewerten, umzusetzen und zu entwickeln. Je nach Rolle und Aufgabenbereich inkludiert dies Fähigkeiten zum sicheren Umgang mit Daten und KI-Systemen im Alltag, über Fähigkeiten zur Identifikation und Bewertung von sicherheitsrelevanten Risiken bei Nutzung und Betrieb von KI-Systemen und zur Umsetzung und Bewertung von Sicherheitsmaßnahmen, bis hin zu Fähigkeiten zur Entwicklung und Verankerung von Sicherheitsstrategien.

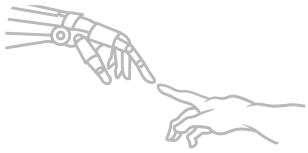
Diese erweiterten Fähigkeiten sind insbesondere für Akteure und Rollen erforderlich, die Verantwortung für die Planung, Ausgestaltung, Umsetzung, Überwachung und Kommunikation von Sicherheitsmaßnahmen tragen – etwa IT-Sicherheitsbeauftragte, KI-Entwickler, Auditoren, Sicherheitsarchitekten oder Datenschutzbeauftragte. Sicherheitskompetenz ergänzt somit die bestehenden Kompetenzarten und befähigt, Sicherheit als integralen Bestandteil des KI-Produktlebenszyklus zu gestalten.

Die nachfolgenden Abschnitte beschreiben einen eigenständigen Kompetenzrahmen für Sicherheitskompetenz und zeigen dessen Nutzung und Anwendung im KI-Kontext.

5.1 Niveaus und Rollenbezug

Mit dem praxisorientierter KI-Kompetenzrahmen aus [12] können Organisationen systematisch die KI-Kompetenzen ihrer Mitarbeiter erfassen, bewerten und dokumentieren. Dieser Kompetenzrahmen

- benennt zehn an KI-Projekten typischerweise beteiligte Akteure und Rollen,
- identifiziert für verschiedene Kompetenzarten drei aufeinander aufbauende Kompetenzniveaus mit zugehörigen Kenntnissen und Fähigkeiten, und
- ordnet jeder Akteurs-Rolle für jede Kompetenzart ein mindestens notwendiges Kompetenzniveau zu.



So können Teams ihre Kompetenzen gezielt weiterentwickeln, um den vielfältigen Anforderungen von KI-Projekten gerecht zu werden. Tabelle 5.1 zeigt eine Übersicht über die Akteurs-Rollen aus [12] mit ihren jeweiligen Aufgaben. Im Folgenden erweitern wir den Kompetenzrahmen aus [12] um die zusätzliche Kompetenzart der Sicherheitskompetenz und unterteilen diese in drei Niveaus.

Tabelle 5.1: Wesentlichen KI-Akteure und ihre Aufgaben.

Akteur	**Aufgaben**
Qualitäts-verantwortliche	Qualitätsstandards Risikomanagement Schulungen
Entscheidende	Strategische Ausrichtung Ressourcenzuteilung
Datenexperten	Datenanalyse Datenmodellierung
KI-Experten	Modellentwicklung Modellanalyse
Compliance-verantwortliche	Regulatorische Anforderungen
Fachexperten	Fachwissen zu Geschäftsdaten und -prozessen
Anwendende	Feedback zur Nutzung
IT-Experten	Technische Entwicklung Betrieb und Support
Koordinatoren	Projektorganisation, -planung und -überwachung
Betroffene	Interessenvertretung (Rechte, Arbeitsbedingungen, Privatsphäre)

Grundlegendes Niveau

Erforderlich für: Koordinatoren, Anwendende, Betroffene

Grundlegendes Niveau ist an den Nutzern ausgerichtet und bezieht sich auf das private und berufliche Umfeld. Es befähigt dazu, sicherheitsrelevante Aspekte im eigenen Umgang mit KI-Systemen zu erkennen, einzuordnen und die eigene digitale Arbeitsumgebung entsprechend sicher zu gestalten. Dazu gehören Fähigkeiten, grundlegende Schutzmaßnahmen wie sichere Passwörter, Zwei-Faktor-Authentifizierung oder Zugriffskontrollen anzuwenden, und einfache Angriffe auf die eigene digitale Arbeitsumgebung wie Phishing-Mails zu erkennen. Dieses Niveau ist vor allem für Anwendende und Koordinatoren relevant, die keine direkte Verantwor-

tung für die sicherheitstechnische Ausgestaltung oder Überwachung von KI-Systemen tragen, aber regelmäßig mit ihnen arbeiten oder sie betreuen.

Weiterführendes Niveau

Erforderlich für: Entscheidende, Datenanalysten, IT-, KI- und Fachexperten, Qualitätsverantwortliche

Weiterführendes Niveau erfordert vertiefte Kenntnisse zur Identifikation, Bewertung, Umsetzung und Kommunikation sicherheitsrelevanter Anforderungen und Maßnahmen sowie Fähigkeiten zur Mitgestaltung der Sicherheit von KI-Systemen in konkreten Anwendungsfällen innerhalb der eigenen Domäne. Dazu gehören Kenntnisse typischer Bedrohungsszenarien in der eigenen Domäne, wie etwa Data Poisoning, Modellmanipulation oder Datenlecks, sowie relevanter rechtlicher, ethischer und technischer Anforderungen an die Sicherheit von KI-Systemen (z.B. DSGVO, ISO/IEC 27001, KI-VO Art. 15). Dieses Niveau umfasst zudem Fähigkeiten zur bedarfsgerechten Umsetzung von Sicherheitsmaßnahmen wie Datenverschlüsselung, Monitoring, Logging und Zugriffskonzepten, sowie zur interdisziplinären Zusammenarbeit. Dieses Niveau ist erforderlich für Akteure, die fachlichen, technischen oder organisatorischen Einfluss auf die Identifikation von Sicherheitsanforderungen und Durchführung Sicherheitsmaßnahmen haben, und sicherheitsbezogene Entscheidungen in Entwicklung oder Betrieb von KI-Systemen treffen. Dieses Niveau wird typischerweise von Prozessverantwortlichen, Projektleitern, Datenanalysten, IT-, KI- und Fachexperten und Datenschutzbeauftragten benötigt.

Expertenniveau

Erforderlich für: Compliance-Verantwortliche, Sicherheitsarchitekten

Expertenniveau beinhaltet die Fähigkeit zur umfassenden Sicherheitsanalyse in komplexen domänenübergreifenden Zusammenhängen, Wahrnehmung übergreifender Sicherheitsmanagement-Führungsrollen, strategischen Steuerung, Entwicklung von Standards und Best Practises sowie zur Schulung und Anleitung anderer. Dieses Niveau ist notwendig für Rollen wie IT-Sicherheitsbeauftragte, Sicherheitsarchitekten, Auditoren und Compliance-Verantwortliche.

Tabelle 5.2 zeigt die Sicherheitskompetenzniveaus mit zugeordneten Rollen. Die Erweiterung des KI-Kompetenzrahmens erfasst Fähigkeiten in Cybersicherheit, Robustheit und Governance.

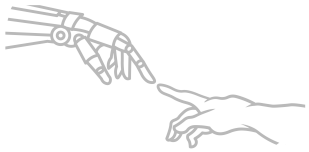


Tabelle 5.2: Sicherheitskompetenz

Sicherheitskompetenz	Notwendige Kenntnisse und Fähigkeiten
Grundlegendes Niveau Betroffene Anwendende Koordinatoren	**Grundkenntnisse im eigenen Anwendungskontext** ► Kenntnis grundlegender Sicherheitsziele wie Vertraulichkeit, Integrität und Verfügbarkeit (CIA-Prinzip) ► Kenntnis grundlegender möglicher Sicherheitslücken und Angriffsmöglichkeiten beim Umgang mit digitalen und KI-Systemen ► Kenntnis einfacher Schutzmaßnahmen beim Umgang mit digitalen und KI-Systemen **Praktische Erfahrung im eigenen Anwendungskontext** ► Fähigkeit, die eigene digitale Arbeitsumgebung sicher zu gestalten ► Fähigkeit, grundlegende Sicherheitsfunktionen beim Umgang mit digitalen und KI-basierten Systemen anzuwenden ► Fähigkeit, einfache Angriffe auf die eigene digitale Arbeitsumgebung zu erkennen **Selbstreflexion** ► Kenntnis der Grenzen der eigenen Sicherheitskompetenz ► Fähigkeit, situationsgerecht geeignete Unterstützung einzuholen
Weiterführendes Niveau Entscheidende IT-Experten Qualitätsverantwortliche KI-Experten Fachexperten	**Weiterführende Kenntnisse in der eigenen Domäne** ► Kenntnis sicherheitsrelevanter Aspekte beim Einsatz von KI-Systemen in spezifischen Geschäftsprozessen oder Fachbereichen ► Kenntnis typischer Bedrohungsszenarien für KI-Systeme in der eigenen Domäne ► Kenntnis relevanter rechtlicher, ethischer und technischer Anforderungen an die Sicherheit von KI-Systemen **Domänenspezifische Mitgestaltung der Sicherheit von KI-Systemen** ► Fähigkeit, potenzielle Angriffsvektoren und Schwachstellen im eigenen Anwendungsbereich zu identifizieren und fachlich zu bewerten ► Fähigkeit, sicherheitsrelevante Maßnahmen im eigenen Anwendungsbereich zu bewerten, bedarfsgerecht umzusetzen und kontinuierlich zu verbessern ► Fähigkeit, sicherheitsbezogene Anforderungen und Ziele für KI-Systeme im eigenen beruflichen Kontext zu definieren ► Fähigkeit, bei Auswahl, Integration und Betrieb von KI-Systemen Sicherheitsstandards einzuhalten **Interdisziplinäre Kommunikation** ► Fähigkeit, sicherheitsbezogene Aspekte von KI-Systemen über verschiedene Fachgebiete hinweg zu kommunizieren und zu diskutieren **Risikomanagement in konkreten Anwendungsfällen** ► Fähigkeit, potenzielle sicherheitsrelevante Risiken im Lebenszyklus von KI-Systemen zu erkennen und im eigenen Handeln zu berücksichtigen
Expertenniveau Compliance- Verantwortliche	**Umfassende Kenntnisse in Theorie und Praxis** ► Umfassende Kenntnis von KI-Sicherheitsrisiken, -anforderungen und -konzepten in verschiedenen Anwendungskontexten ► Fähigkeit, Sicherheitsrisiken, -anforderungen und -maßnahmen am KI-System umfassend zu erfassen, zu analysieren und zu bewerten ► Fähigkeit, innovative Sicherheitsarchitekturen und Schutzmechanismen für KI-Systeme zu entwickeln und umzusetzen ► Fähigkeit, sicherheitsrelevante Auswirkungen neuer Technologien zu bewerten und entsprechend zu handeln **Führungsrolle und Strategie** ► Fähigkeit, die sicherheitsbezogene Ausgestaltung von KI-Systemen fachlich zu leiten und zu überwachen ► Fähigkeit, übergreifende strategische Sicherheitsentscheidungen in KI-Projekten zu treffen ► Fähigkeit, organisationale Sicherheitsstrategien für KI-Systeme zu entwickeln, umzusetzen und zu verantworten ► Fähigkeit, Sicherheitsbedarf der Organisationsebene mit Chancen und Risiken der KI-Nutzung zu verknüpfen und strategisch zu bewerten ► Fähigkeit, neue gesetzliche, technologische und ethische Entwicklungen im KI-Sicherheitsbereich zu antizipieren und deren Relevanz einzuordnen **Wissenstransfer** ► Fähigkeit, KI-Sicherheitsrisiken, -anforderungen und -konzepte auf neue Anwendungskontexte zu übertragen und anzupassen ► Fähigkeit, KI-Sicherheitskompetenz zu vermitteln und bei der sicheren Entwicklung und Nutzung von KI-Systemen zu beraten und anzuleiten ► Fähigkeit, sicherheitsbezogene Leitlinien, Standards und Best Practices für Entwicklung und verantwortungsvollen Einsatz von KI-Systemen zu entwickeln

5.2 Nutzung und Anwendungsszenarien

Die erfassten Kompetenzen lassen sich, wie in [12] beschrieben, mittels Kompetenzspinnen visualisieren. Eine Kompetenzspinne bildet dabei die Ausprägung verschiedener Kompetenzarten auf einer vierstufigen Skala (0 bis 3) entlang radialer Achsen ab. Hierbei entspricht Stufe 3 dem Expertenniveau, Stufe 2 dem weiterführenden Niveau, Stufe 1 dem grundlegenden Niveau und Stufe 0 dem Fehlen von Kompetenz in einer Kompetenzart. So lassen sich individuelle und teambezogene Kompetenzprofile anschaulich und vergleichbar darstellen.

Die Sicherheitskompetenzspinnen kommen insbesondere in drei Szenarien zum Einsatz:

- **Individuelle Kompetenzprofile** helfen Mitarbeitenden bei der Selbstverortung, etwa für Rollenzuweisungen in sicherheitsrelevanten Projekten oder als Nachweis im Rahmen von Audits.
- **Soll-Ist-Abgleiche** ermöglichen den Abgleich individueller Kompetenzen mit angestrebten Zielprofilen, um Qualifizierungsbedarfe systematisch zu erkennen.
- **Teamprofile** unterstützen die kompetenzbasierte Zusammenstellung interdisziplinärer Teams für Entwicklung, Prüfung oder Überwachung sicherheitskritischer KI-Systeme.

Die Visualisierung schafft Transparenz über vorhandene und erforderliche Kompetenzen – sowohl für einzelne Personen als auch für ganze Teams.

Mitarbeitende können ihre Kompetenzen selbst einschätzen und in Form einer Spinne darstellen. Dies erleichtert den Vergleich mit einem Soll-Profil, das beispielsweise die Anforderungen einer spezifischen Rolle abbildet. Abweichungen werden so auf einen Blick erkennbar (siehe Abbildung 5.1) und können Grundlage für individuelle Weiterbildungsmaßnahmen oder Nachweise im Rahmen von Audits sein.

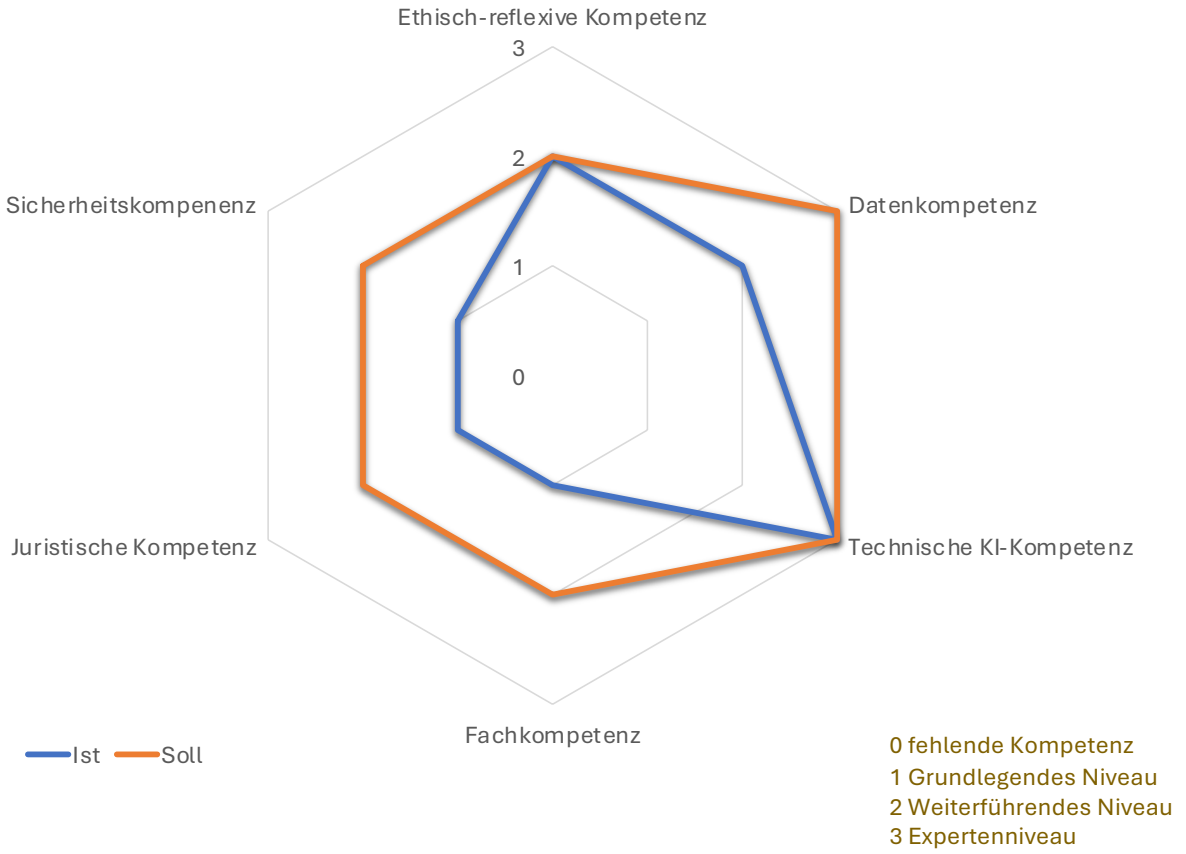
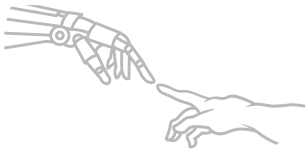


Abbildung 5.1: Exemplarischer Soll-Ist-Vergleich einer Person.

Für interdisziplinäre Teams, etwa in Cybersicherheitsaudits oder Notfallstrukturen, kann auf Basis individueller Spinnen ein kollektives Teamprofil erstellt werden. Die Überlagerung der Einzelprofile zeigt, ob alle relevanten Kompetenzarten angemessen vertreten sind oder ob gezielte Qualifizierungen oder zusätzliche Ressourcen nötig sind. Dies unterstützt die strategische Personalplanung und fördert eine effektive Zusammenarbeit zwischen unterschiedlichen Fachbereichen.

Insgesamt unterstützt die Sicherheitskompetenzspinne sowohl die strategische Personal- und Teamplanung als auch die operative Qualifizierung im Rahmen der Sicherheitsstrategie für KI-Systeme. Sie liefert eine nachvollziehbare, prüfbare Dokumentationsbasis für interne Zwecke und für externe Prüfinstanzen – insbesondere im Rahmen der technischen Dokumentation nach Artikel 15 der KI-Verordnung.



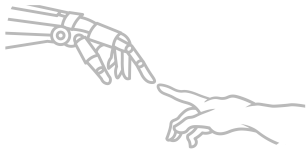
6 Zusammenfassung

Die Gewährleistung der Cybersicherheit in KI-Systemen ist ein fundamentaler Bestandteil, um die Konformität mit den Anforderungen der EU-KI-Verordnung, insbesondere Artikel 15, sicherzustellen. Sie dient als Nachweis dafür, dass ein KI-System während seines gesamten Lebenszyklus sicher, robust und zuverlässig funktioniert.

Der vorliegende Leitfaden zeigt, wie diese Anforderungen strukturiert und praktisch umgesetzt werden können, um die Entwicklung und den Betrieb verantwortungsvoller KI-Systeme zu gewährleisten. KI-Systeme sind dabei spezifischen und komplexen Bedrohungsszenarien ausgesetzt, die über die klassische IT-Sicherheit hinausgehen.

Zu den zentralen Angriffen zählen *Data Poisoning*, bei dem Trainingsdaten manipuliert werden, um die Modellleistung zu beeinträchtigen, *Model Poisoning*, also die gezielte Veränderung von Modellparametern zur Integration von Hintertüren, *Adversarial Examples*, die das Modell durch kaum wahrnehmbare Eingaben zu Fehlklassifikationen verleiten, *Model Extraction*, bei dem Angreifer versuchen, das Modell durch systematische Abfragen zu rekonstruieren, sowie *Data Leakage*, also die unbeabsichtigte Offenlegung sensibler Daten. Hinzu kommen klassische Angriffe auf Vertraulichkeit, Integrität und Verfügbarkeit, die im KI-Kontext besondere Schutzmaßnahmen erfordern.

Um diesen Risiken zu begegnen, muss Cybersicherheit ganzheitlich über den gesamten Produktlebenszyklus eines KI-Systems integriert werden – von der Konzeption bis zur Abschaltung. Bereits in der Konzeptionsphase sind Bedrohungsanalysen, Risikobewertungen, die Integration von *Security by Design*-Prinzipien sowie die Berücksichtigung von Supply-Chain-Security entscheidend.



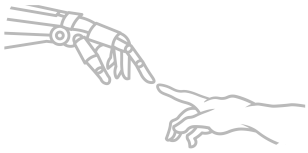
In der Datenphase stehen eine rechtssichere Datenerhebung unter Einhaltung der DSGVO, die Sicherstellung von Datenqualität und Manipulationsschutz durch Validierung, Anomalieerkennung und Herkunftsnachweise sowie robuste technisch-organisatorische Maßnahmen wie Zugriffskontrollen und Verschlüsselung im Vordergrund.

In der Modellierungsphase sind die Absicherung der Trainingsprozesse, die Nachvollziehbarkeit durch Versionierung, die Anwendung robuster Trainingsverfahren wie Adversarial Training oder Defensive Distillation und der Schutz gegen Modell-Diebstahl oder Model Extraction Attacks zentrale Aufgaben. Auch lernverfahrensspezifische Risiken wie Reward Hacking oder Angriffe im Federated Learning erfordern gezielte Schutzmechanismen. Vor der Produktivsetzung gilt es, Infrastruktur und Schnittstellen zu härten, Zugriffskonzepte anzupassen und die Integrität aller Systembestandteile durch Security-Assessments wie Fuzzy Testing und semantisches Testen abzusichern.

Während des Betriebs spielen kontinuierliches Monitoring, Validierung von Eingaben und Ausgaben, strikte Zugangskontrollen sowie ein geregeltes Update- Management eine zentrale Rolle, um Angriffen wie Evasion-Attacks oder Prompt Injection zu begegnen.

In der Abschaltungsphase können ebenfalls Risiken bestehen, sodass die sichere Löschung von Daten und Modellartefakten, der Entzug aller Zugriffsrechte und eine vollständige Dokumentation unverzichtbar sind.

Über die technischen Maßnahmen hinaus ist eine umfassende Governance-Struktur notwendig. Dazu gehören die Definition klarer Zuständigkeiten, etwa durch die Rolle eines KI-Security-Officers, die Etablierung konsistenter Richtlinien für Data Governance und sicheres Modelltraining sowie die kontinuierliche Neubewertung der Bedrohungslage. Ergänzend ist der Aufbau umfassender Sicherheitskompetenzen innerhalb der Organisation unerlässlich. Ein dreistufiger Kompetenzrahmen



von grundlegenden über weiterführende bis hin zu Expertenkenntnissen ermöglicht es, die Fähigkeiten der Mitarbeitenden systematisch zu erfassen und zu entwickeln – von Anwendern bis zu Sicherheitsarchitekten und Compliance-Verantwortlichen. Dies unterstützt sowohl die individuelle Weiterentwicklung als auch die strategische Teamplanung und trägt wesentlich dazu bei, den komplexen Anforderungen an die KI-Sicherheit gerecht zu werden.

Literaturverzeichnis

[1] Das Europäische Parlament und der Rat der Europäischen Union. *EU Artificial Intelligence Act*. 2024. URL: https://artificialintelligenceact.eu/de/das-gesetz/ (siehe S. 1, 7, 52).

[2] BSI. *Künstliche Intelligenz*. 2025. URL: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html (siehe S. 9).

[3] NIST. *Adversarial Machine Learning*. 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf (siehe S. 11).

[4] Verordnung (EU) 2016/679. *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR)*. 2016 (siehe S. 17, 20).

[5] ISO/IEC. *ISO/IEC 27000:2018*. 2018. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:27000:ed-5:v1:en. (siehe S. 17).

[6] Andrea Apicella, Francesco Isgrò und Roberto Prevete. *Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning*. 2024. URL: https://arxiv.org/abs/2401.13796 (siehe S. 21).

[7] Jiaxin Fan u. a. "A survey on data poisoning attacks and defenses". In: *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*. IEEE. 2022, S. 48–55 (siehe S. 21).

[8] Ibrahim M Ahmed und Manar Younis Kashmoola. "Threats on machine learning technique by data poisoning attack: A survey". In: *International Conference on Advances in Cyber Security*. Springer. 2021, S. 586–600 (siehe S. 21).

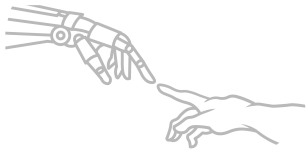
[9] NIST. *Identity & Access Management*. besucht April 2025. URL: https://www.nist.gov/identity-access-management (siehe S. 23).

[10] Nastaran Farhadighalati u. a. "A Systematic Review of Access Control Models: Background, Existing Research, and Challenges". In: *IEEE Access* 13 (2025), S. 17777–17806. DOI: 10.1109/ACCESS.2025.3533145 (siehe S. 23).

[11] NIST. *Post-Quantum Cryptography*. besucht April 2025. URL: https://csrc.nist.gov/pqc-standardization (siehe S. 24).

[12] Alina Lorenz u. a. *KI-Kompetenzrahmen für Organisationen - Ein praktisches Instrument zur Umsetzung der EU-KI-Verordnung*. 2025 (siehe S. 43, 52–54, 57).

Abbildungsverzeichnis



Tabellenverzeichnis

Universität Augsburg University

Audit- und Wissensplattform für Künstliche Intelligenz