APP.4.10 Sicherheit von KI-Systemen

1. Beschreibung

1.1. Einleitung

Systeme der künstlichen Intelligenz (KI) und des maschinellen Lernens (ML) werden zunehmend zu integralen Bestandteilen von Geschäftsprozessen und kritischen Anwendungen. Sie unterscheiden sich fundamental von traditioneller Software: Ihr Verhalten wird nicht explizit programmiert, sondern durch den Lernprozess aus Daten abgeleitet. Dieser datenzentrierte Ansatz schafft eine neue Klasse von Angriffsvektoren, die den gesamten Lebenszyklus eines KI-Systems betreffen – von der Erhebung der Trainingsdaten über den Trainingsprozess bis hin zum produktiven Betrieb des gelernten Modells (Inferenz). Die Sicherheit von KI-Systemen umfasst daher nicht nur die Absicherung der zugrundeliegenden Infrastruktur, sondern insbesondere den Schutz der Daten, der Algorithmen und der Modelle selbst. Bedrohungen wie Datenvergiftung, die gezielte Umgehung von Modellen durch "Adversarial Attacks" oder der Diebstahl von geistigem Eigentum in Form des Modells erfordern spezifische, an die KI-Domäne angepasste Schutzmaßnahmen.

1.2. Zielsetzung

Dieser Baustein zeigt einen systematischen Weg auf, um die spezifischen Risiken im Lebenszyklus von KI- und ML-Systemen zu adressieren. Ziel ist es, die Vertraulichkeit, Integrität und Verfügbarkeit von KI-Modellen und deren Datenbasis sicherzustellen. Es sollen Maßnahmen etabliert werden, um die Systeme widerstandsfähiger gegen KI-spezifische Angriffe zu machen, die Nachvollziehbarkeit von KI-Entscheidungen zu fördern und eine sichere Integration in den Informationsverbund zu gewährleisten.

1.3. Abgrenzung und Modellierung

Der Baustein APP.4.10 ist auf alle Systeme anzuwenden, in denen KI- oder ML-Modelle entwickelt, trainiert oder betrieben werden.
Er behandelt nicht die allgemeinen Sicherheitsanforderungen an die zugrundeliegende IT-Infrastruktur (siehe SYS-Bausteine) oder die generelle sichere Softwareentwicklung (siehe CON.1). Die sichere Speicherung der Daten wird durch die Bausteine APP.4.3 bis APP.4.9 ergänzt. Dieser Baustein fokussiert auf die spezifischen Sicherheitsaspekte des MLOps-Lebenszyklus. Ethische und datenschutzrechtliche Fragestellungen (z. B. Fairness, Bias) werden hier nicht primär behandelt, jedoch unterstützen die hier definierten Sicherheitsmaßnahmen deren Einhaltung.

2. Gefährdungslage

Für den Baustein APP.4.10 sind folgende spezifische Bedrohungen und Schwachstellen von besonderer Bedeutung:

Datenvergiftung (Data Poisoning)

Angreifer manipulieren die Trainingsdaten, indem sie fehlerhafte, böswillige oder irreführende Daten einschleusen. Dies führt dazu, dass das KI-Modell falsche Muster lernt, systematische Fehler macht oder eine "Hintertür" enthält, die der Angreifer später ausnutzen kann.

Gezielte Umgehung und Manipulation von Modellen (Adversarial Evasion)

Angreifer erstellen zur Laufzeit speziell präparierte Eingabedaten, die für Menschen unauffällig sind, das KI-Modell aber zu einer gezielten Fehlentscheidung verleiten. Beispiele sind geringfügig veränderte Bilder, die falsch klassifiziert werden, oder Spam-E-Mails, die als legitim erkannt werden.

Diebstahl von KI-Modellen und Trainingsdaten (Model Inversion/Extraction)

Durch wiederholte, gezielte Abfragen an die API eines KI-Modells kann ein Angreifer dessen internes Verhalten rekonstruieren und ein funktional äquivalentes Modell stehlen (Model Extraction). In manchen Fällen können sogar sensible Informationen aus den ursprünglichen Trainingsdaten wiederhergestellt werden (Model Inversion).

Offenlegung von Trainingsdaten durch das Modell (Data Leakage)

Das KI-Modell "erinnert" sich an spezifische, oft seltene Beispiele aus seinen Trainingsdaten (Overfitting). Bei bestimmten Eingaben kann es diese sensiblen Informationen, z. B. personenbezogene Daten, im produktiven Betrieb preisgeben.

Unsichere Integration und Orchestrierung der ML-Pipeline

Die gesamte Kette von Werkzeugen zur Erstellung und Bereitstellung von KI (MLOps-Pipeline) ist ein attraktives Ziel. Eine Kompromittierung des Quellcode-Repositorys, des Artefakt-Speichers für Modelle oder der Deployment-Skripte kann zur Manipulation des gesamten KI-Systems führen.

3. Anforderungen

Im Folgenden sind die spezifischen Anforderungen des Bausteins APP.4.10 aufgeführt.

3.1. Basis-Anforderungen

Die folgenden Anforderungen MÜSSEN für diesen Baustein vorrangig erfüllt werden.

APP.4.10.A1 Richtlinie für den sicheren Einsatz von KI (B)

Es MUSS eine Richtlinie existieren, die die Ziele, Verantwortlichkeiten und grundlegenden Sicherheitsregeln für die Entwicklung und den Betrieb von KI-Systemen in der Institution festlegt.

APP.4.10.A2 Inventarisierung und Klassifizierung von Trainingsdaten (B)

Alle für das Training von KI-Modellen verwendeten Datensätze MÜSSEN inventarisiert und entsprechend ihrer Schutzbedürftigkeit klassifiziert werden. Der Zugriff auf diese Daten MUSS geregelt sein.

APP.4.10.A3 Grundlegende Absicherung der Modell-Schnittstellen (API) (B)

Der Zugriff auf KI-Modelle, die über APIs bereitgestellt werden, MUSS authentifiziert und autorisiert werden. Die Schnittstellen MÜSSEN gegen gängige Web-Angriffe geschützt sein.

APP.4.10.A4 Protokollierung von Trainingsläufen und API-Anfragen (B)

Alle Trainingsläufe MÜSSEN protokolliert werden, um nachvollziehen zu können, welches Modell mit welchen Daten und Parametern trainiert wurde. Anfragen an die produktiven Modell-APIs MÜSSEN ebenfalls protokolliert werden.

3.2. Standard-Anforderungen

Gemeinsam mit den Basis-Anforderungen entsprechen die folgenden Anforderungen dem Stand der Technik. Sie SOLLTEN grundsätzlich erfüllt werden.

APP.4.10.A5 Sicherstellung der Integrität von Trainingsdaten (S)

Es SOLLTEN technische Maßnahmen ergriffen werden, um die Integrität der Trainingsdaten zu schützen. Dies umfasst die Überprüfung der Datenquellen, die Nutzung von Checksummen und die Absicherung der Datenspeicher gegen unautorisierte Veränderungen.

APP.4.10.A6 Absicherung der ML-Pipeline (S)

Der gesamte Lebenszyklus eines Modells von der Entwicklung über das Training bis zum Deployment SOLLTE durch eine abgesicherte MLOps-Pipeline verwaltet werden. Dies umfasst die Absicherung von Code-Repositories, Trainingsumgebungen und Artefakt-Speichern.

APP.4.10.A7 Implementierung von grundlegenden Abwehrmaßnahmen gegen Adversarial Attacks (S)

Es SOLLTEN robuste Programmierpraktiken und Techniken zur Validierung und Bereinigung von Eingabedaten eingesetzt werden, um einfache Adversarial Attacks zu erschweren.

APP.4.10.A8 Versionierung und Herkunftsverfolgung von Daten und Modellen (S)

Es SOLLTE ein System zur Versionierung von Datensätzen und Modellen etabliert sein. Es MUSS jederzeit nachvollziehbar sein, welche Version eines Modells mit welcher Datenversion trainiert wurde (Data and Model Provenance).

APP.4.10.A9 Überwachung des Modellverhaltens im produktiven Betrieb (S)

Die Vorhersagen und Entscheidungen des KI-Modells im produktiven Betrieb SOLLTEN überwacht werden, um Abweichungen vom erwarteten Verhalten (Model Drift) oder Anzeichen für Angriffe frühzeitig zu erkennen.

3.3. Anforderungen bei erhöhtem Schutzbedarf

Die folgenden Anforderungen sind exemplarische Vorschläge für ein Schutzniveau, das über den Stand der Technik hinausgeht. Sie SOLLTEN bei erhöhtem Schutzbedarf in Betracht gezogen werden.

APP.4.10.A10 Implementierung von Schutzmaßnahmen gegen Modell-Extraktion (H)

Der Zugriff auf die Modell-API SOLLTE überwacht werden, um verdächtige Abfragemuster zu erkennen, die auf einen Extraktionsangriff hindeuten. Maßnahmen wie Rate-Limiting oder das Hinzufügen von "Wasserzeichen" zu den Modellausgaben SOLLTEN in Betracht gezogen werden.

APP.4.10.A11 Durchführung von Adversarial-Robustness-Tests (H)

KI-Modelle SOLLTEN proaktiv mit spezialisierten Frameworks auf ihre Widerstandsfähigkeit gegen Adversarial Attacks getestet werden. Die Ergebnisse SOLLTEN genutzt werden, um die Modelle durch robuste Trainingstechniken (Adversarial Training) zu härten.

APP.4.10.A12 Einsatz von Techniken zur Erklärbarkeit und Interpretierbarkeit (Explainable AI, XAI) (H)

Es SOLLTEN Techniken aus dem Bereich der Explainable AI eingesetzt werden, um die Entscheidungen des Modells nachvollziehbar zu machen. Dies hilft, versteckte Fehler, Manipulationen und Datenlecks zu identifizieren.

APP.4.10.A13 Nutzung von Privacy-Enhancing Technologies im Trainingsprozess (H)

Beim Training mit sensiblen, insbesondere personenbezogenen Daten SOLLTEN Techniken wie Differential Privacy oder Federated Learning eingesetzt werden, um den Schutz der Privatsphäre in den Trainingsprozess zu integrieren und das Risiko von Data Leakage zu minimieren.