Leitfaden KI & Informations-sicherheit

Ein Überblick zu Informationssicherheit von und durch KI

bitkom

Inhalt

1 Einleitung

1.1 Zielstellung

Der Fachkräftemangel in Deutschland verschärft sich jährlich. Allein in der IT-Branche waren 2022 ganze 137.000 IT-Stellen in der Gesamtwirtschaft unbesetzt und 2023 steigerte sich die Zahl der unbesetzten IT-Stellen auf 149.000[1]. Neben langfristigen politischen Lösungsansätzen wie Schulbildung im Bereich Informatik und Einwanderung von Fachkräften bieten auch technologische Lösungen wie der Einsatz von Künstlicher Intelligenz (KI) die Möglichkeit, die unternehmenseigene Effizienz und Effektivität signifikant zu steigern und den Fachkräftemangel zumindest teilweise auszugleichen. Trotz dieser Chancen nutzen lediglich 20 % der Unternehmen KI[2]. Dies liegt auch daran, dass Unternehmen sich um neue unbekannte Risiken sorgen, welche sich aus dem Einsatz von KI-Modellen und KI-Systemen ergeben könnten. Insgesamt betrachten dabei 2023 69 % der Unternehmen neue IT-Sicherheitsrisiken als eine Gefahr im Kontext des KI-Einsatzes[3].

69 %

der Unternehmen sehen neue IT-Sicherheitsrisiken als Gefahr beim Einsatz von KI. (Bitkom, 2023)

Dieser Leitfaden bietet einen strukturierten Überblick über das Zusammenspiel von Informationssicherheit und KI. Er adressiert zunächst grundlegende Aspekte der **Informationssicherheit von KI**, einschließlich der Sicherheit von KI-Modellen und KI-Systemen und möglicher Angriffsszenarien und deren Eintrittswahrscheinlichkeit. Darauf aufbauend werden **Schutzmaßnahmen für KI-Systeme** vorgestellt, die zu einer resilienten und sicheren Nutzung von KI im Unternehmenskontext beitragen. Im zweiten Teil zeigt der Leitfaden auf, wie **KI zur Stärkung der IT-Sicherheit** selbst beitragen kann, etwa durch *Security Event Monitoring* und *Threat Detection*, *Identity & Access Management*, *Endpoint Protection* sowie *Data Leakage Prevention*.

1.2 Was ist Künstliche Intelligenz?

Unter dem Begriff KI verstehen wir eine Technologie, die durch mehrere Ansätze und Techniken, z. B. die des »Maschinellen Lernens«, schwer zu formalisierende Probleme wie die Bildklassifizierung oder Sprachverarbeitung lösen kann. Hierbei zeigt die Technologie Fähigkeiten, die in beschränktem Maße mit menschlicher Intelligenz assoziiert werden können. In einem Teilgebiet der KI, dem Maschinellen Lernen, wird das Verhalten eines KI-Modells mittels Algorithmen aus zumeist großen Datenmengen angelernt. Hierbei erzielen sie bei vielen Aufgaben bessere Ergebnisse als herkömmliche Verfahren, die nur auf klar definierten und fest programmierten Regelwerken basieren. Die dazu zählenden Algorithmen lernen beispielsweise, aus Daten Muster zu erkennen oder gewünschte Verhaltensweisen zu zeigen, ohne dass jeder Einzelfall explizit programmiert wurde. Werden für die KI-Modelle »große« mehrschichtige neuronale Netze verwendet, spricht man in diesem Fall vom »Deep Learning«.

Definition KI-System nach KI-VO Art.3 (1): Ein KI-System ist :„ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können.“

[1] Rekord-Fachkräftemangel: In Deutschland sind 149.000 IT-Jobs unbesetzt | Presseinformation | Bitkom e.V.

[2] Künstliche Intelligenz in Deutschland - Perspektiven aus Bevölkerung und Unternehmen (2024) | Bitkom e.V.

[3] Deutsche Wirtschaft drückt bei Künstlicher Intelligenz aufs Tempo (2023) | Presseinformation | Bitkom e.V.

In den folgenden Kapiteln fokussieren wir uns auf die weitverbreiteten KI-Systeme und Modelle basierend auf Maschinellem Lernen und stellen dar, welche neuartigen Angriffsvektoren existieren. Diese sollten bei der Einführung von KI im Unternehmen berücksichtigt werden. Zunächst wollen wir aber noch kurz auf die Unterscheidung von KI-Systemen und KI-Modellen eingehen:

Unterscheidung KI-System und KI-Modell

Im Folgenden wollen wir uns für die Begriffsunterscheidung an der KI-Verordnung der Europäischen Union (KI-VO) orientieren, die am 1. August 2024 in Kraft getreten ist. Bei einem KI-System handelt es sich um ein umfassenderes maschinengestütztes System, das neben dem KI-Modell selbst auch Infrastruktur, Datenquellen, Schnittstellen (APIs), Benutzerinteraktionen und Entscheidungslogiken beinhalten kann. Der Begriff KI-Modell hingegen erfasst das technische Herzstück hinter einem KI-System. Letzteres meint – am Beispiel eines Künstlichen Neuronalen Netzes (KNN) – bspw. die Architektur und Anzahl der Neuronen und Schichten sowie die dahinterstehenden Algorithmen und Gewichtungen. Diese Unterscheidung ist beispielsweise für die KI-Verordnung (KI-VO) besonders relevant. Die KI-VO reguliert sowohl KI-Systeme (nach der nebenstehenden Definition im roten Kasten) als auch speziell im Fall der sogenannten »KI-Modelle mit allgemeinem Verwendungszweck« (General-Purpose-AI Modelle, kurz GPAI-Modelle), die Technologie hinter dem KI-System.

Definition des GPAI-Modells in Art. 3 Nr. 63 der KI-VO:
Ein KI-Modell mit allgemeinem Verwendungszweck ist „ein KI-Modell — einschließlich der Fälle, in denen ein solches KI-Modell mit einer großen Datenmenge unter umfassender Selbstüberwachung trainiert wird —, das eine erhebliche allgemeine Verwendbarkeit aufweist und in der Lage ist, unabhängig von der Art und Weise seines Inverkehrbringens ein breites Spektrum unterschiedlicher Aufgaben kompetent zu erfüllen, und das in eine Vielzahl nachgelagerter Systeme oder Anwendungen integriert

2 Informationssicherheit von KI

KI bringt im Business-Umfeld große Stärken im Sinne vielfältigster Anwendungsmöglichkeiten mit sich. Wer als Unternehmen in bestimmten Prozessen KI verantwortungsvoll einsetzen möchte, muss sich aber bewusst sein, dass mit der neuen Technologie auch Risiken einhergehen, die über bereits bekannte Angriffsflächen und Verwundbarkeiten aus dem Bereich der klassischen IT-Sicherheit hinausgehen. Auf die KI-spezifischen Verwundbarkeiten wird daher im folgenden Abschnitt eingegangen. Es sei darauf hingewiesen, dass hier nur die wichtigsten KI-spezifischen Angriffsarten dargestellt werden können. Darüber hinaus gibt es noch eine Vielzahl speziellerer Angriffe, die allerdings oftmals Varianten der Grundangriffsarten darstellen und daher nicht Teil der weiterführenden Betrachtungen dieses Dokuments sind.

»Die größte Schwäche der künstlichen Intelligenz ist ihre größte Stärke: Sie weiß nichts außerhalb dessen, was wir ihr beibringen.« – Garry Kasparow

Im Wesentlichen existieren zwei KI-spezifische »neue« Angriffsfelder: Einerseits können die Algorithmen oder Modelle, insbesondere die für Maschinelles Lernen geeigneten neuronalen Netze, angegriffen werden. Andererseits sind die Daten, die für Training, Test und Betrieb von KI-Systemen benötigt werden, mögliche Schwachstellen. Bevor diese genauer erläutert werden, soll aber auf die Grundlagen der IT-Sicherheit von KI-Modellen und -Systemen eingegangen werden, welche die Voraussetzung für einen sicheren Betrieb darstellen.

2.1 Grundlagen der KI-Sicherheit

Datenqualität = KI-Qualität

Die Qualität und Performanz, mit der KI-Modelle und -Systeme klassifizieren, prognostizieren und Inhalte generieren, stehen und fallen mit der Qualität der zum Testen, Trainieren und im operativen Betrieb verwendeten Daten. Das aggregierte, antrainierte »Wissen« aus riesigen Datenmengen und dessen Anwendung ist eine einzigartige Stärke von KI, die gleichzeitig jedoch angesichts der großen Datenabhängigkeit neue Verwundbarkeiten mit sich bringt (siehe z. B. Kapitel 2.3 »Data Poisoning«). Daher sollte sichergestellt werden, dass die vielfältigen Möglichkeiten von KI unter bestmöglicher Absicherung datenspezifischer Schwachstellen genutzt werden können, um im Unternehmenskontext beispielsweise zur Automatisierung bestimmter Aufgaben und damit zur Realisierung vielschichtiger unternehmensbezogener Ziele und Wettbewerbsvorteile beizutragen. Damit wird die Datenqualität bei der Nutzung von KI zur zentralen Determinante und ist über den gesamten Lebenszyklus der Daten zu betrachten. Dieser umfasst die Erhebung, Verarbeitung und Vorbereitung der Daten, damit diese von einem KI-Modell genutzt werden können. In diesem Zusammenhang gilt, dass die eingesetzten Systeme nur so gut sein können wie ihre Grundlagen.

Wie später zu lesen ist, gibt es auch KI-spezifische Angriffsmöglichkeiten, die nur auf Datensätze abzielen. Dabei ist zu beachten, dass diese Datensätze zunehmend aus verschiedenen Datenarten, die für das Training von KI eingesetzt werden, bestehen. Die wichtigsten Datenmodi sind derzeit Bilder (einzeln und bewegt), Audiodateien, Zeitreihendaten und Texte. Für unterschiedliche KI-Anwendungen sind unterschiedliche Datenarten notwendig.

Sicherheit cloudbasierter KI-Dienste

KI-Systeme und KI-Modelle auf Basis der verschiedenen Datentypen müssen dabei nicht zwingend von denjenigen Unternehmen selbst entwickelt werden, die sie später intern nutzen oder Dritten zur Nutzung als Dienstleistung anbieten wollen. Es existieren auch Anbieter, die KI-~~Systeme~~ über sogenannte Software-as-a-Service-Schnittstellen (SaaS) anbieten, oder das cloudbasierte Hosting eigenentwickelter oder nachtrainierter KI-Modelle und KI-Systeme ermöglichen.

Durch die Nutzung derartiger Schnittstellen bei entsprechenden Anbietern findet gegebenenfalls die Verlagerung eines Teils der KI-Schwachstellen, beispielsweise mit Blick auf den direkten Modellzugriff hin zum Cloudanbieter, statt. Verwundbarkeiten hingegen, welche die dem Modell gestellten Anfragen (Inferenz) betreffen, müssen gegebenenfalls weiterhin vom ursprünglichen Unternehmen berücksichtigt werden. Daneben erfordern klassische Cloud-Verwundbarkeiten Berücksichtigung, die bereits im Rahmen zahlreicher Schutzmaßnahmen, Richtlinien und Zertifikate adressiert werden. Beispielhaft sei hier der *Cloud Computing Compliance Criteria Catalogue* (C5) des BSI genannt, welcher durch den *Artificial Intelligence Cloud Service Compliance Criteria Catalogue* (AIC4) um KI-spezifische Aspekte erweitert wurde. Zu beachten ist außerdem,

dass sich Unternehmen bei der Nutzung vollständig vorgefertigter Cloud-KI-Services in Abhängigkeiten zum betreffenden Dienstleister hinsichtlich der verwendeten Daten begeben können. Beispielsweise sind die vom Dienstleister zum Training verwendeten Daten gegebenenfalls nicht einzusehen oder veränderbar. Damit ist ebenfalls nicht gewährleistet, dass der gewählte Anbieter das KI-Modell und/oder das KI-System gegen die in Kapitel 2.3. beschriebenen Angriffsarten abgesichert hat.

2.2 Varianten von KI-Modellen

Neben den oben genannten Angriffsmöglichkeiten auf die Daten spielen das eigentliche Modell und dessen zugrundeliegende Architektur und Arbeitsweise eine wichtige Rolle für die Sicherheit.

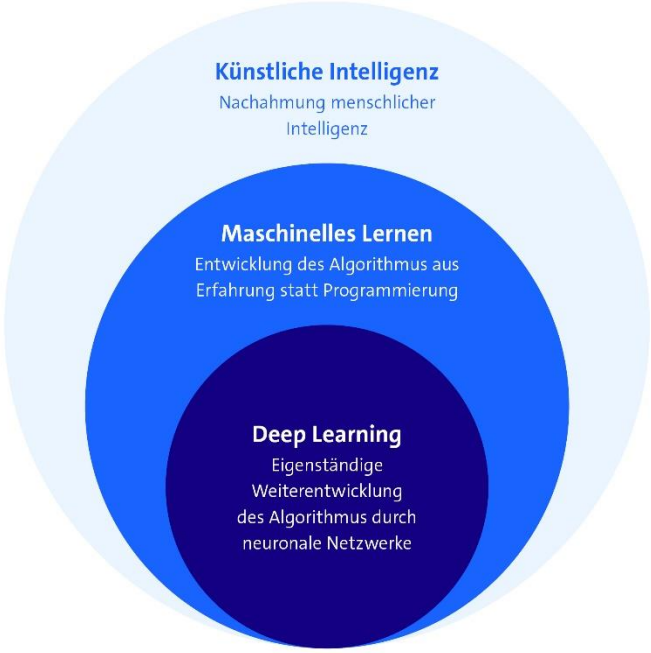


Abbildung 1: Übersicht des Verhältnisses von KI, Maschinellem Lernen und Deep Learning.

Modelle, die über die Fähigkeit des Maschinellen Lernens verfügen und ihr Verhalten durch Training ändern, stellen aus der Perspektive der Informationssicherheit und auch aus Safety-Aspekten zusätzliche Herausforderungen dar, die spezielle Security-Maßnahmen erfordern. Um Angriffe durch passende Maßnahmen möglichst zu verhindern, ist es wichtig, das verwendete Modell zu kennen sowie dessen Arbeitsweise zu verstehen oder nachvollziehen zu können. Bestmögliche Erklärbarkeit (Explainability; XAI) des Modells und der Arbeitsweise ist für die Bewertung der Sicherheit elementar, aber bei sogenannten Grey- oder Blackbox-Systemen schwer oder nur indirekt möglich.

Es gibt eine Vielzahl von Modellen für Maschinelles Lernen. Ihre Auswahl hängt von der Art der Anwendung oder Funktion ab, die umgesetzt werden soll, sowie der Anzahl der verfügbaren bzw. zu verarbeitenden oder notwendigen Trainingsdaten. Dabei gewinnen Architekturen mit vielfältigen Anwendungsmöglichkeiten wie sogenannte *General*

Purpose AI[4] an Bedeutung gegenüber *Narrow AI* (»schwacher« KI), die sich auf eine begrenzte Aufgabe konzentriert.

Methoden des Maschinellen Lernens können auch unterschieden werden in

- **Statistische Methoden,** z. B. verschiedene Regressionsmodelle, oder Modellvarianten unter Verwendung von Entscheidungsbäumen, oder Modelle, die zur Klassifizierung oder Clusterbildung geeignet sind. Diese Modelle können meist erklärbar gestaltet werden.
- **Deep Learning (DL)** mit Methoden wie beispielsweise Deep Neural Networks (DNNs) und Convolutional Neural Networks (CNNs) für komplexe Aufgaben wie Bilderkennung und natürliche Sprachverarbeitung (Natural Language Processing; NLP).
- **Transformer-Modelle**, die eine Weiterentwicklung der DNN-Architektur sind, die durch die Verarbeitung von Sequenzen bei Bildern, Texten und Sprache weitere Leistungsverbesserung ermöglicht haben.
- **Generative KI-Modelle,** eine weitere DNN-Variante, die selbstständig aus einer Eingabefrage neue Varianten auf Basis gelernter Trainingsdaten generieren und die beispielsweise in Verbindung mit **Large Language Models** (sehr große Sprachmodelle; LLM) beeindruckende Ergebnisse erzielen. Prominente LLMs sind:
  - GPT-Serie (Generative Pre-trained Transformer) von OpenAI
  - BERT (Bidirectional Encoder Representations from Transformers) von Google
  - ERNIE (Enhanced Representation through Knowledge Integration) von Baidu
- **Diffusion-Modelle,** die die Idee der Diffusion nutzen, um zum Beispiel aus eingegebenen Texten neue realistische und vielfältige Bilder zu erzeugen. Beispiele sind Dall-E 2 von OpenAI, sowie Stable Diffusion, ein Open-Source-Modell von Stability AI.

Allgemein sind Deep Learning-Modelle und neuronale Netze Teil hochkomplexer, selbstveränderlicher Systeme. Sie werden als Grey- oder Blackbox-Systeme bezeichnet, da die Regeln, aus denen das Ergebnis ermittelt wurde, nicht mehr transparent, erklärbar und nachvollziehbar sind.

Diese Zusammenstellung ist nicht abschließend. In jedem Fall haben die Auswahl des passenden KI-Modells sowie dessen Anwendungskontext großen Einfluss auf die Sicherheit des gesamten KI-Systems.

2.3 Angriffe auf KI

Angriffe auf KI-Systeme umfassen die Manipulation der Trainingsdaten (»Data Poisoning«), die zielgerichtete Nutzung der Eingabedaten für einen Angriff auf das KI-Modell (»Model Evasion«) und die Gewinnung von Informationen über das Modell (»Model Extraction«) während der Trainings- oder Betriebsphase.

[4] Siehe u.a. die General Purpose AI-Definition in der KI-VO.

Data Poisoning

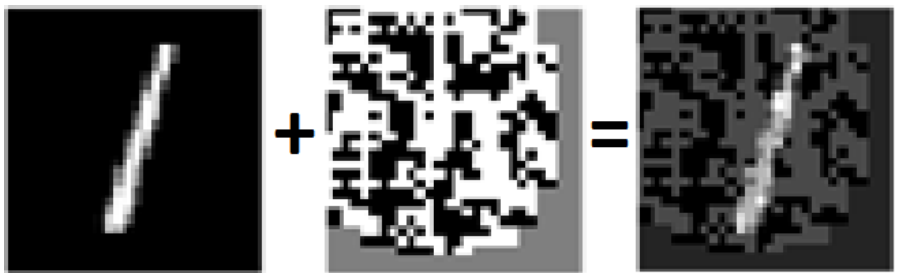
Abbildung 2: Bild aus dem CIFAR-100 Datensatz mit eingebautem Muster

»Data Poisoning« erzeugt eine Fehlfunktion oder Leistungsminderung von Maschinellen Lernmodellen. Eine Variante ist die Manipulation des Trainingsdatensatzes durch den Angreifer. Ein einfacher »Poisoning«-Angriff besteht darin, in den Trainingsdaten die Klasse einer Eingabe auf die gewünschte Klasse umzuschreiben (z. B. soll ein »Vogel« aus dem CIFAR-100-Datensatz[5] nicht mehr als »Vogel«, sondern als »Katze« klassifiziert werden). Anschließend werden die manipulierten Daten in den Trainingsdatensatz injiziert und während des Modelltrainings verwendet. Diese Angriffsmethode kann auch bei einer nur kleinen Anzahl manipulierter Datensätze wirksam sein. Die manipulierten Trainingsdaten sind unter Umständen durch manuelle Prüfung oder andere Verfahren erkennbar. In der Regel ist eine Detektion jedoch herausfordernd.

Eine weitere Variante der »Poisoning«-Angriffe ist der »Backdoor«-Angriff, der darauf abzielt, ein Modell so zu manipulieren, dass eine bestimmte Reaktion auf ein Muster in einer Eingabe erfolgt. Bei der Bilderkennung können diese Muster die Form von kleinen Artefakten oder schwer erkennbaren Wasserzeichen in den Eingangsdaten annehmen (siehe Abbildung 2). Das Muster wird in eine Teilmenge der Trainingsdaten eingebaut und durch den Angreifer mit der gewünschten Klasse versehen. Die Grundidee ist, dass das Modell während des Trainings mit den kompromittierten Trainingsdaten lernt, das Muster mit der vom Angreifer festgelegten, falschen Klasse zu assoziieren. Die Erfolgsrate von »Backdoor«-Angriffen hängt von der Architektur des Modells, der Anzahl kompromittierter Trainingsbeispiele und der Beschaffenheit der gewählten Auslösemuster ab. Ein Muster mit einer hohen Erfolgsrate beeinflusst nicht unbedingt die Klassifikationsleistung des Modells bei gutartigen Eingaben. Daher sind »Backdoor«-Modelle schwer zu erkennen. Bei der Verwendung von bereits vortrainierten Modellen aus öffentlichen Quellen ist unbedingt zu beachten, dass auch Risiken wie eingebaute Hintertüren übernommen werden könnten.

Model Evasion

Bei einem »Model Evasion«-Angriff versucht der Angreifer, während der Abfrage (statt während des Trainings) einer KI eine Fehlklassifizierung zu verursachen. Der Angreifer konstruiert eine manipulierte Eingabe, indem er seiner Eingabe eine kleine Störung hinzufügt. Im Fall von Bildern ist diese u. U. für den Menschen kaum wahrnehmbar. Die so erzeugten Eingaben werden auch als *adversariale Beispiele* bezeichnet. Es kann



[5] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images

Abbildung 3: Adversariales Beispiel aus dem MNIST-Datensatz. Das Addieren der Störung bewirkt eine Klassifikation der Zahl 1 als 4.

zwischen gezielten Angriffen, bei denen der Angreifer das Modell zwingt, die gewünschte Klasse vorherzusagen, und ungezielten Angriffen, die eine allgemeine Fehlklassifizierung oder Verringerung der Prognosegüte verursachen, unterschieden werden. Die Angriffe können in der physischen oder digitalen Welt stattfinden. Bestimmte Muster könnten beispielsweise dazu führen, dass ein autonom fahrendes Auto Verkehrszeichen verwechselt oder ein biometrisches Kamerasystem die Identität einer Person falsch zuordnet.

Model Extraction

Extraktionsangriffe fassen alle Angriffe zusammen, die darauf abzielen, das Modell oder Informationen aus den Trainingsdaten zu rekonstruieren. Für den Angreifer hilfreich sind bei solchen Angriffen Vorkenntnisse über den Trainingsdatensatz des Modells oder Zugriff auf öffentlich verfügbare Teile davon.

Für Unternehmen, die erhebliche Ressourcen in die Entwicklung eines KI-Modells investiert haben, stellt der Diebstahl ihres Modells eine Bedrohung der Geschäftsgrundlage dar. So können Angreifer neben dem klassischen Eindringen in die IT-Infrastruktur das Modell durch zahlreiche Abfragen, deren Antworten sie in ihr eigenes Schattenmodell einspeichern, kopieren. Der Diebstahl des Modells kann als Sprungbrett für weitere Angriffe dienen.

Zur Klasse der Extraktionsangriffe zählt der »Membership Inference«-Angriff, bei dem der Angreifer versucht festzustellen, ob ein bestimmter Datensatz Teil der Trainingsmenge des Modells war. Das Vorhandensein einer bestimmten Person in einer Trainingsmenge kann eine sensitive Information sein. Bei »Attribute Inference«-Angriffen versucht der Angreifer, die Vertraulichkeit der Trainingsdaten des Modells zu verletzen, indem er den Wert eines sensiblen Attributs ermittelt, das einer bestimmten Person oder Identität in den Trainingsdaten zugeordnet ist.

Die Rekonstruktion exemplarischer Klassen aus den Trainingsdaten durch einen Angreifer wird als »Model Inversion«-Angriff bezeichnet. Ein mögliches Angriffsszenario könnte darin bestehen, das Gesicht einer Person allein durch Zugriff auf die Ausgaben einer für deren Erkennung trainierten KI zu rekonstruieren.

Risiken von Large Language Models (LLMs)

Wie bereits erwähnt, sind große KI-Sprachmodelle (eng.: Large Language Models, kurz: LLMs) statistische Modelle aus der Familie der Generativen KI. Für eine Texteingabe (Prompt) erzeugen sie eine Textausgabe, die eine wahrscheinliche Fortsetzung des Texts darstellt. Um diese Fähigkeit zu erlernen, werden die Modelle üblicherweise auf großen Textkorpora mit diversen Texten trainiert. Den meisten LLMs liegt dabei die Transformer-Architektur zugrunde, die effizient große Mengen an Textdaten analysieren und mittels Aufmerksamkeitsmechanismen Zusammenhänge in der Sprache erler-

nen kann. Anwendungsmöglichkeiten von LLMs beinhalten die automatisierte Generierung, Übersetzung und Zusammenfassung von Text sowie die Interaktion als Chatbot mit Usern auf Basis von hinterlegten Textinhalten.

Die vielfältigen Anwendungsmöglichkeiten von LLMs sind jedoch auch mit Risiken verbunden. Zu den allgemeinen Risiken zählen die Reproduktion von Fehlinformationen, Vorurteilen und diskriminierenden Aussagen durch das Training eines LLMs auf fehlerbehafteten oder einseitigen Datensätzen. Des Weiteren ist aufgrund des probabilistischen Charakters der Modelle nicht garantiert, dass der erzeugte Text faktisch korrekt ist. Das potenzielle Erfinden von Inhalten wird auch als »Halluzinieren« bezeichnet. Bei der Nutzung eines externen LLMs können die Eingaben unter Umständen vom Anbieter des Modells weiterverwendet werden. Je nach Nutzungsbedingungen können diese Daten durch den Anbieter auch für das eigene Training verwendet werden und sind damit potenziell durch Extraktionsangriffe gefährdet. Hier besteht für Geschäftsgeheimnisse und persönliche Daten ein Risiko. Bei der Verwendung von LLMs, deren Trainingsprozess nicht bekannt ist, ist ebenfalls das Risiko eingebauter Backdoors zu beachten.

Das für Nutzerinnen und Nutzer unbemerkte Einbringen von Eingaben in ein LLM durch einen Angreifer wird als »Indirect Prompt Injection«[6] bezeichnet. Dieser Angriff betrifft LLMs, die auf Daten aus externen Quellen zugreifen oder ungeprüfte Dokumente Dritter als Input verwenden. Der Angreifer verbirgt dabei auf Webseiten oder in Dokumenten Anweisungen an das LLM. Dies kann dazu führen, dass ein autonomer Agent mit einem LLM zur Anweisungsverarbeitung einem Angreifer Zugriff auf Daten oder Systeme verschafft. Bei einem LLM als Chatbot könnten beispielsweise versteckte Anweisungen dazu führen, dass dieser die Nutzerinnen und Nutzer im Gesprächsverlauf dazu bewegt, auf einen Link zu klicken. Durch das Anklicken des Links könnten vertrauliche Informationen aus dem Gespräch abgegriffen werden.

2.4 Eintrittswahrscheinlichkeiten von Angriffen auf KI und Handlungsbedarfe

Zum gegenwärtigen Zeitpunkt fallen dokumentierte Angriffe auf KI-Schwachstellen – im Gegensatz zu klassischen Cyberangriffen (z. B. DDoS, Datenlecks, Botnets) – kaum ins Gewicht.[7] Eine der ersten empirischen Studien zu Angriffen auf KI kann zudem kein systematisch erhöhtes Gefahrenpotenzial für einzelne der untersuchten Wirtschaftszweige, darunter Gesundheitswesen, IT-Sicherheit und industrielle Fertigung, identifizieren. Insgesamt berichteten ca. 17 % der 139 befragten industriellen Anwenderinnen und Anwender von Täuschungs- und Umgehungsversuchen mit Blick auf die verwendeten KI-Modelle und -Systeme. Davon konnten drei als Evasion-Attacken im Personalwesen respektive der Bilderkennung beim autonomen Fahren identifiziert werden, während zwei Poisoning-Angriffe mit beispielsweise bewusstem Falsch-Labeling von Trainingsdaten durch Mitarbeiterinnen und Mitarbeiter beschrieben wurden.[8] Unabhängig vom konkreten Anwendungskontext lassen sich jedoch bestimmte systemspezifische Faktoren identifizieren, die eine gesteigerte Exposition gegenüber einzelnen

[6] Indirect Prompt Injections - Intrinsische Schwachstelle in anwendungsintegrierten KI-Sprachmodellen I Bundesamt für Sicherheit in der Informationstechnik
[7] Artificial Intelligence Incident Database.
[8] Grosse et.al.: Machine Learning Security in Industry, 2023

Angriffsarten mit sich bringen: So begünstigt unter anderem ein via API öffentlich zugänglicher KI-Service Evasion-Attacken, während Trainingsdatensätze unbekannter Herkunft, unbeschränkter Modellzugang und kontinuierliches Weiterlernen (»Online Machine Learning«) den Erfolg von Data Poisoning-Attacken wahrscheinlicher machen.[9]

Für die Zukunft ist mit einer deutlichen Zunahme von Angriffen auf KI-spezifische Schwachstellen auszugehen. Zum Beispiel nutzen Safety-Anwendungen wie Zugangssysteme in Zukunft verstärkt eine KI-basierte Gesichtserkennung und Zuordnung der Zutrittsberechtigung. Durch eine nicht ordnungsgemäß abgesicherte Trainingsdaten-Lieferkette kann ein KI-Modell z. B. durch einen Data Poisoning-Angriff kompromittiert werden und berechtigten Mitarbeiterinnen und Mitarbeitern den Zutritt verweigern oder unberechtigten Personen den Zutritt erlauben. Ähnliche Schwachstellen können zur Manipulation und Sabotage zum Beispiel autonomer Fahrsysteme und klinischer KI-Systeme ausgenutzt werden.

Eine Echtzeit-Manipulation ohne zwangsläufiges Data Poisoning im Vorfeld ist beispielsweise bei KI-Systemen denkbar, die auf Basis von Zeitreihendaten die autonome Fertigungssteuerung unterstützen. Insbesondere bei der Integration von Sensordaten aus unsicheren oder unbekannten Domänen lassen sich so während der Laufzeit über erfolgreiche Model Evasion-Angriffe möglicherweise Fehlfunktionen etablieren, die den Produktionsprozess erheblich stören und etwa Roboter falsche Handlungen durchführen lassen.

Große Sprachmodelle sowie darauf basierende Chat-Anwendungen, die sich für automatisierte Texterkennung und -generierung eignen, werden von einigen Unternehmen bereits zur automatisierten Bewerberfilterung eingesetzt. Hier würde ein erfolgreicher Model-Extraction-Angriff auf ein beispielsweise anhand von realen Bewerbungen nachtrainiertes Modell möglicherweise sensible Bewerberdaten offenlegen und damit erhebliche datenschutzrechtliche Konsequenzen sowie massiven Reputationsverlust für das betroffene Unternehmen nach sich ziehen.

Im Angesicht der vielfältigen Möglichkeiten und tatsächlich beobachteten – bislang geringen – Häufigkeiten sollten die genannten Angriffsszenarien allerdings auch unter dem Gesichtspunkt der ökonomischen Vorteilhaftigkeit für den potenziellen Angreifer betrachtet werden. Alle drei vorgestellten Angriffsarten müssen nicht nur wohldurchdacht und vorbereitet sein, sondern erfordern auch signifikanten Zeitaufwand, Zugriffsmöglichkeiten auf Input- und Output-Datenströme (oder gar das Modell selbst) und nicht zuletzt erhebliche Expertise in der Funktionsweise Künstlicher Intelligenz. »Klassische« (cyber-)kriminelle Methoden, sich physischen oder informationstechnischen Zugang zu Betriebsräumen und Systemen zu verschaffen, sind zwar auffälliger. In den meisten Fällen sind sie derzeit aber noch mit wohl erheblich geringeren Ressourcenaufwänden für Angreifer verbunden. Gleichwohl sollten die KI-spezifischen Gefährdungen für kritische Anwendungsfälle im Rahmen einer Risikoanalyse systematisch bewertet werden.

Die KI-VO verleiht dieser Notwendigkeit in Zukunft auch regulatorischen Nachdruck. KI-Systeme, welche nach der KI-VO unter die Hochrisiko-Stufe fallen, müssen u. a. ein

[9] Bundesamt für Sicherheit in der Informationstechnik: Security of AI-Systems: Fundamentals - Adversarial Deep Learning, 2022.

angemessenes Maß an »Genauigkeit, Robustheit und Cybersicherheit« erreichen und dies während ihres Lebenszyklus aufrechterhalten. Zur Spezifizierung dieser Anforderung treibt die EU-Kommission zusammen mit relevanten Interessengruppen und Organisationen die Entwicklung von Benchmarks und Messmethoden voran.

Weiterhin werden GPAI-Modelle, wie bereits oben erwähnt, separat reguliert. Diese werden noch einmal unterteilt in »einfache« Modelle und solche Modelle mit »systemischen Risiken«. Für GPAI-Modelle mit systemischen Risiken gelten zusätzliche Sicherheitsanforderungen, welche beispielsweise das Einrichten eines Risikomanagementsystems oder Red Teaming-Tests umfassen. Auch wird hier die explizite Anforderung eines angemessenen Maßes an Cybersicherheit für das Modell gestellt. Konkret soll ein Modell demnach etwa vor allem vor unbeabsichtigten Modelllecks, nicht genehmigten Veröffentlichungen, der Umgehung von Sicherheitsmaßnahmen, unbefugten Zugriffen und Modell-Diebstahl geschützt werden. Die Anforderungen und Wege zur Erfüllung dieser sollen über sogenannte »Codes of Practice« (Praxisleitfäden) spezifiziert werden, solange es keine harmonisierten Standards gibt.

Nicht zuletzt sollte angesichts der bevorstehenden Herausforderungen die derzeit noch bestehende »Ruhephase« genutzt werden, um die Absicherung von KI-Systemen im Kontext der jeweiligen Anwendungsszenarien auch im Bereich der angewandten Grundlagenforschung deutlich auszubauen.

2.5 Schutzmaßnahmen für KI-Systeme

Sieht man zunächst von den Besonderheiten einer Künstlichen Intelligenz bezogen auf Daten und algorithmische Modelle ab, handelt es sich aus Security-Perspektive um eine Softwarelösung in Verbindung mit einer Infrastruktur. Hier kommen die bisher bewährten Security-Maßnahmen, wie der Betrieb eines Information-Security-Managementsystems (ISMS), mit üblichen technischen und organisatorischen Themen nach anerkanntem Stand der Technik zum Einsatz. Hierzu gehört z. B. eine klassische Risikobetrachtung.

KI-spezifische IT-Sicherheits-Themen umfassen einerseits die Überwachung und den Schutz der (großen) Datenmengen. Andererseits sind die Eignung, Resilienz und Robustheit der KI-Modelle für die Sicherheit im geplanten Einsatz von großer Bedeutung und eine neue Herausforderung.

Sicherheit von Daten
Die Herausforderungen zur Erkennung von gefälschten Daten sind kein neues Thema. Neben den konventionellen Schutz- und Sicherheitsmaßnahmen der IT-Sicherheit sollten beim Einsatz von KI deutlich mehr Aufmerksamkeit und Schutzmaßnahmen auf die Datenherkunft sowie Datenmanipulation (Data Poisoning) und Datenverlust gelegt werden. Die Rückverfolgbarkeit von Daten ist aktuell eine der großen Herausforderungen und eine der Quellen von Sicherheitsproblemen. Daten, aber auch andere Komponenten können sehr komplexe Lebenszyklen haben, aus vielen Quellen stammen, transformiert und erweitert werden und von Drittanbietern inklusive Open Source stammen.

Integrität und Qualität bzw. Eignung sind entscheidend für die Sicherheit. Unter Datenintegrität versteht man die Konsistenz, die Richtigkeit, die Vertrauenswürdigkeit und die Rekonstruierbarkeit der Daten während ihrer gesamten Lebensdauer. Dies umfasst Maßnahmen, die zum Ziel haben, dass geschützte Daten während der Beschaffung, Verarbeitung oder Übertragung nicht durch unautorisierte Personen entfernt oder verändert werden können.

Die Datenqualität selbst ist der Bestandteil eines Daten-Managements, welches geeignete Prozesse zur Beurteilung und Bereinigung der Daten definiert und die Fragen beantwortet:

- Ist die Herkunft der Daten sowie deren Authentizität und Integrität, Lebenszyklus-Zertifikate, End-to-End-Datenherkunft (»Data Provenance«) identifizierbar?
- Wie ist die Form der Veredlung (Definieren, Sammeln, Selektieren, Umwandeln, Verifizieren) und Anreicherung der Rohdaten zu Modell- oder Trainingsdaten?
- Gibt es menschliche Beteiligung (z. B. User-Feedback oder Labelling) an den Entscheidungsfindungen innerhalb einer Verarbeitung?
- Sind die Institutionen, die die Daten für das KI-System liefern, transparent und vertrauenswürdig?
- Sind die Daten auf geeignete Weise kuratiert?
- Ist der Einbau von Prüfankern und Prüfagenten vorgesehen?

Unterstützen können unter anderem sogenannte Reputationssysteme, algorithmische Analysen der Ausgabe und deren Einschränkungen sowie Authentizitätsanalysen zur Überprüfung der konkreten Herkunft der Daten.

Sicherheit von KI-Modellen
Ein KI-ML-Modell kann einerseits durch die gezielte Verwendung (Angriff) von manipulierten Daten verändert werden und andererseits durch eine direkte Störung bzw. einen Angriff auf die Modellarchitektur beispielsweise durch Veränderung oder Manipulation des neuronalen Netzes oder der Betriebsbedingungen wie zu wenig verfügbaren Speicherplatzes.

Ein robustes, resilientes Modell sollte in der Lage sein, zuverlässige und konsistente Vorhersagen unter verschiedensten Bedingungen zu erzeugen, am besten einschließlich solcher, die im Trainingsprozess nicht berücksichtigt wurden. Die Robustheit eines Modells ist abhängig von der spezifischen Anwendungsdomäne und muss für den jeweiligen Anwendungsfall definiert, getestet, sowie im laufenden Betrieb immer wieder überprüft werden. Die »Accuracy« (Genauigkeit) der Ergebnisse, d. h. wie gut die richtigen Vorhersagen im Vergleich zu allen Vorhersagen insgesamt ausfallen oder ungewollter »Bias« (Diskriminierung) zu beobachten ist, sind weitere Kriterien.

Verbesserte und größtmögliche Robustheit und Sicherheit können beispielsweise folgende Konzepte bieten:

Adversarial Training:

Adversarial Training ist eine Technik im maschinellen Lernen, die verwendet wird, um die Robustheit und Sicherheit von Modellen gegenüber feindlichen Angriffen zu erhöhen. Bei dieser Technik werden absichtlich manipulierte Daten, sogenannte »adversariale« Beispiele, in das Training eingebracht, um das Modell darauf vorzubereiten, auf potenziell schädliche Eingaben zu reagieren. Eine Methode des Adversarial Trainings ist die »Adversarial Defence through Randomization« (ADR). Diese arbeitet mit stochastischen Elementen, die in das Modell integriert sind, um es gegenüber Angriffen widerstandsfähiger zu machen.

DARTS (Differentiable Architecture Search):
DARTS stellt einen Ansatz für die automatische Architektursuche dar, der versucht, Modelle zu finden, die besser generalisieren und robuster gegenüber Veränderungen in den Eingabedaten sind.

Certified Robustness:
Diese Modelle bieten nachweislich Robustheit und mathematische Garantien für die Robustheit gegenüber bestimmten Arten von Störungen.

Zusammenfassend kann gesagt werden, dass nicht ausreichend umgesetzte konventionelle IT-Sicherheit, unpassend ausgewählte Modelle sowie mangelnde Robustheit und Genauigkeit zu Sicherheitsrisiken führen. Um diese beurteilen zu können, ist eine weitgehende Transparenz, Erklärbarkeit und Nachvollziehbarkeit erforderlich.
Die Forschung in diesen Bereichen ist ein aktives Gebiet, und es werden laufend neue Ansätze entwickelt, um die Zuverlässigkeit und Sicherheit von KI-Systemen zu verbessern. Für die ganzheitliche Betrachtung der mit KI verbundenen Sicherheitsrisiken und damit deren effektiver Reduzierung müssen Unternehmen Praktiken, Richtlinien und Kontrollen entwickeln. Im Folgenden finden Sie einige Empfehlungen, die man dazu ansetzen kann:

- Führen Sie regelmäßig eine umfassende Sicherheitsbewertung der aktuellen KI-Systeme des Unternehmens durch, welche auch die KI-spezifischen Risiken berücksichtigt
  - Identifizierung potenzieller Schwachstellen für Angriffe
  - Bewertung der Risiken, z. B. durch Simulation von Angriffen
  - Erstellung eines Berichts über die Ergebnisse
  - Ableitung von technischen oder organisatorischen Maßnahmen sowie Prüfung der Wirksamkeit, z. B. durch Simulation von Angriffen
- Entwicklung einer Sicherheitsrichtlinie und sicherer Verfahren für den Einsatz von KI
  - Definition von Richtlinien für die Erkennung, Meldung und Reaktion auf Angriffe
  - Bewertung und Berücksichtigung der für das KI-System relevanten Compliance- und Regulierungsaspekte
  - Entwicklung eines Schulungsprogramms für Mitarbeiterinnen und Mitarbeiter zu bewährten Sicherheitspraktiken und potenziellen Risiken
  - Ableitung von Kriterien für die Beschaffung von KI-Systemen und benötigten Komponenten (z. B. Daten)
- Implementierung strenger Zugangskontrollen und Authentifizierungsmaßnahmen
  - Entwickeln und Implementieren von Zugangskontrollen, um den unbefugten Zugriff auf Modelle und Daten des Maschinellen Lernens zu verhindern

  - Testen und validieren Sie die Wirksamkeit der Zugangskontrollen
  - Protokollieren Sie nach Möglichkeit den Zugriff auf und die Veränderung von Daten und Modellen
- Investieren Sie in Technologien und Techniken zur Verteidigung gegen Angriffe
  - Identifikation und Bewertung von Techniken zur Abwehr von KI-Angriffen
  - Auswahl und Bereitstellung geeigneter Techniken für die KI-Systeme des Unternehmens
- Regelmäßige Überwachung und Prüfung von KI-Systemen
  - Einen regelmäßigen Zeitplan für die Überwachung und Prüfung von KI-Systemen aufstellen
  - Entwicklung und Umsetzung von Verfahren zur Erkennung von und Reaktion auf Angriffe
- Etablierung eines Notfallplans für den Fall eines Angriffs auf KI-Systeme
  - Definition klarer Rollen und Verantwortlichkeiten im Krisenfall
  - Erstellung eines Kommunikationsplans zur schnellen Information aller relevanten Verantwortlichen im Notfall

3 KI für die IT-Sicherheit

Sicherheitsteams stehen heute vor der Herausforderung, nicht nur externe Bedrohungen zu überwachen und aufzuspüren. Sie sind zusätzlich mit technischen Werkzeugen konfrontiert, die für die Verarbeitung großer Datenmengen in der Cloud – insbesondere unter Einsatz von Künstlicher Intelligenz – nicht entwickelt wurden.
Die Angriffsflächen von Organisationen sind in den vergangenen Jahren immer größer geworden und generieren immer mehr Sicherheitsdaten, sowohl quantitativ als auch qualitativ. Netzwerk-, Endpunkt-, Identitäts- und Cloud-Daten verbleiben dabei vielerorts in getrennten Systemen. Die Endpunkt-Telemetrie sitzt meistens isoliert in einem »Endpoint Detection and Response« (EDR)-System. Cloud-Daten befinden sich zudem in einer separaten Cloud-Sicherheitslösung, wobei nur ein Bruchteil von Protokollen, dafür aber eine Flut von Warnmeldungen an das »Security Information and Event Management« (SIEM) gesendet werden. Infolgedessen müssen Security Operations-Analysten Daten manuell analysieren, um Warnungen zu selektieren und wirksame Maßnahmen zu ergreifen. Die Vielzahl an Warnungen überlasten die Analysierenden, sodass Bedrohungen übersehen werden und lange Zeiträume bis zur Erkennung entstehen. Künstliche Intelligenz kann hier auf vielfältige Weise zur Unterstützung genutzt werden.

3.1 Security Event Monitoring und Threat Detection

Der richtige Einsatz Maschinellen Lernens erleichtert die schnelle Identifizierung bekannter bösartiger Verhaltensweisen. So können Profile verbundener Geräte erstellt,

Informationen über regelmäßig erfolgte Aktivitäten gewonnen und dadurch Erkenntnisse darüber generiert werden, was »normal« ist und was nicht. Dies ermöglicht die automatische Anomalie-Erkennung als Komponente des Security Event Monitorings.

Zusätzlich ermöglicht KI in diesem Kontext eine Zero-Day-Erkennung. Bei herkömmlichen Sicherheitslösungen muss eine fehlerhafte Aktion mindestens einmal erkannt werden, damit sie als »maliziöse Aktivität« identifiziert werden kann. Auf diese Weise funktioniert die signaturbasierte Malware-Erkennung älterer Generationen. Maschinelles Lernen kann auf intelligente Weise bisher unbekannte Formen von Malware und Angriffen durch das Korrelieren verschiedener Datenquellen (einschließlich verhaltensbasierter Daten) identifizieren, um dadurch Organisationen vor potenziellen Zero-Day-Angriffen zu schützen.

Insgesamt ermöglicht KI einen Erkenntnisgewinn in großem Umfang im Sinne des Security Event Monitorings und der Threat Detection. Dadurch, dass Daten und Anwendungen immer häufiger an vielen verschiedenen Standorten verwendet werden, ist es schier unmöglich, Entwicklungen und sich manifestierende Trends über eine große Anzahl von Geräten hinweg manuell zu erkennen. Maschinelles Lernen kann dabei unterstützen und so die Generierung von umfangreichen automatisierten Erkenntnissen möglich machen.

3.2 Identity & Access Management

On-Premise-Systeme, Cloud-Dienste oder hybride Lösungen und der durch COVID19 stark geprägte Wandel hin zu einer flexiblen und standortunabhängigen Arbeit über Notebooks, Smartphones oder Tablets stellen Betreiber von IT-Infrastrukturen vor enorme Herausforderungen, digitale Werte ihrer Systemlandschaft adäquat vor illegitimem Zugriff zu schützen. Hierzu ist ein effektives »Identity und Access Management« (IAM) unerlässlich und gesetzlich in Teilen für KMUs vorgeschrieben (vgl. DSGVO, Sarbanes-Oxley Act). Mittels einer zentralen Verwaltung hat IAM zum Ziel, Identitäten (u. a. interne Nutzende, Partner sowie Hard- und Software) nur Zugriff auf diejenigen Informationen und Systeme zu gewähren, für die sie berechtigt sind. Dies gilt über den gesamten Lebenszyklus einer solchen Identität innerhalb der IT-Infrastruktur. Neben dieser Autorisierung ist die Authentifizierung ebenfalls Teil des IAM.

Für die erfolgreiche Etablierung von IAM sind Prozesse und Werkzeuge gleichermaßen von Bedeutung. Hierbei kann Technologie basierend auf KI nicht nur das Sicherheitsniveau erhöhen, sondern auch zu einer Reduzierung von Anmeldenachweisen beitragen. Ein Beispiel ist die Authentifizierung einer Identität über biometrische Merkmale (u. a. Fingerabdruck oder Gesicht), bei der mittels KI neu erfasste biometrische Daten mit Referenzdaten abgeglichen werden, um die behauptete Identität zu bestätigen oder zu widerlegen. Ein weiteres Beispiel ist die sogenannte adaptive Authentifizierung. Über die Zeit erlernt und überwacht eine KI das Verhalten einer Identität. So kann kontextbasiert über Anmeldemuster (u. a. Gerät, Lokalität, Zeiten, Daten) eine Risikobewertung erfolgen und Authentifizierungsanforderungen abhängig vom konkreten Risiko ggf. vermindert oder erhöht werden. Über solche erstellten Profile der KI können gleichzeitig Abweichungen zur Norm ermittelt und potenzielle Sicherheitsverstöße

frühzeitig unterbunden werden (vgl. Anomalie-Erkennung unter Security Event Monitoring und Threat Detection. In diesem Zusammenhang kann KI auch helfen, existente IAM-Strukturen zu optimieren, indem auf Grundlage typischer Zugriffsmuster Empfehlungen für eine verbesserte und schlankere Strukturierung der Rollenorganisation oder von Prozessen ausgesprochen werden. Über konventionelle Datenanalysen ist dies nur schwer erreichbar.

3.3 Endpoint Protection

»Endpoint Protection« bezieht sich auf eine Reihe von Sicherheitsmaßnahmen zum Schutz von Endgeräten. Darunter fallen Laptops, Desktops, mobile Geräte und Server. Das Ziel von Endpoint Protection ist der Schutz der Geräte vor Cyber-Bedrohungen wie Ransomware, Viren, Würmern, Trojanern und anderer bösartiger Software, welche die Sicherheit des Endgeräts und/oder des Netzwerks gefährden.

KI kann den Schutz von Endgeräten erheblich verbessern, indem sie ihn proaktiver, effektiver und effizienter macht. So können KI-Algorithmen beispielsweise Muster und Anomalien im Endpunktverhalten erkennen, potenzielle Schwachstellen identifizieren und Angriffe verhindern. Außerdem können KI-basierte Schutzlösungen aus früheren Angriffen lernen und diese Informationen nutzen, um ihre Erkennungs- und Präventionsfunktionen zu verbessern. Ferner können KI-Algorithmen die Analyse des Netzwerkverkehrs, von Aktivitätsprotokollen und von *Threat Intelligence Feeds* unterstützen, um potenzielle Sicherheitsrisiken zu erkennen und schnell darauf zu reagieren. KI-basierte Endpoint Protection-Lösungen können mithilfe von Algorithmen des Maschinellen Lernens das Benutzerverhalten, den Netzwerkverkehr und die Anwendungsaktivitäten überwachen, um Sicherheitsverletzungen zu erkennen und Gegenmaßnahmen einzuleiten.

Der Endpunktschutz kann durch KI-Algorithmen also wesentlich verbessert werden, indem die Erkennung und Verhinderung von Sicherheitsbedrohungen automatisiert, die Arbeitsbelastung von Sicherheitsteams reduziert und die Effektivität und Effizienz von Sicherheitsmaßnahmen erhöht werden. Da sich Cyber-Bedrohungen ständig weiterentwickeln, werden KI-basierte Lösungen zum Schutz von Endgeräten für Unternehmen immer wichtiger.

3.4 Data Leakage Prevention

Der Begriff »Data Leakage Prevention« (DLP) bezieht sich auf Sicherheitsmaßnahmen zur Verhinderung von unbefugter Übertragung oder Offenlegung von sensiblen Daten. Dabei wird mit DLP das Ziel verfolgt, Datenabfluss durch versehentlich entstandene oder absichtlich herbeigeführte Lecks zu schützen, unabhängig davon, ob diese durch E-Mails, Cloud-Speicher, USB-Laufwerke oder andere Kanäle entstehen.

DLP-Lösungen können durch KI-Algorithmen erheblich verbessert werden. So können die Algorithmen große Datenmengen analysieren, darunter Netzwerkverkehr, Benutzeraktivitätsprotokolle und Datenklassifizierungs-Tags, um potenzielle Datenlecks in Echtzeit zu identifizieren. Hier spielt die Anpassungsfähigkeit durch den Einsatz von KI ebenfalls eine erhebliche Rolle, um die Erkennungs- und Präventionsfunktionalität von DLP-Lösungen stetig zu verbessern.

Auch die Erkennung potenzieller Anomalien durch Analyse des Benutzerverhaltens wird durch den Einsatz von KI möglich. So können beispielsweise abnormale Datenzugriffsmuster, ungewöhnliche Dateiübertragungen oder unbefugte Zugriffsversuche von legitimen Aktivitäten unterschieden werden. Durch die Analyse von Mustern und Trends bei der Datenübertragung können KI-basierte DLP-Lösungen Datenlecks vorhersagen und verhindern, bevor sie auftreten.

Die Nutzung von KI-basierter Sprachanalyse eignet sich zudem zur Analyse von E-Mails und weiteren Kommunikationskanälen, um Hinweise auf Datenlecks zu erkennen, selbst wenn die Daten verschleiert oder verschlüsselt sind. Ferner kann der Datenzugriff auf Grundlage von Benutzeridentität, -rolle und -standort überwacht und gesteuert werden, sodass sensible Daten besser geschützt werden.

4 Fazit und Ausblick

Der Einsatz von KI im Unternehmenskontext bietet sowohl große Chancen als auch neue Risiken, die es zu berücksichtigen gilt. Das Verhalten eines KI-Modells wird wesentlich durch die Trainingsdaten bestimmt, mit denen es trainiert wurde. Damit sind Angriffe auf die Datengrundlage ein möglicher Angriffsvektor, mit dem sich Unternehmen beschäftigen sollten. Eine Manipulation der Trainings- und Eingabedaten kann beispielsweise zu einer Fehlklassifikation führen. Eine weitere KI-spezifische Angriffsmöglichkeit sind Angriffe auf das Modell bzw. den Algorithmus selbst. Diese Bedrohungen sollten bereits beim Design beziehungsweise bei der Implementierung Berücksichtigung finden, um potenzielle Risiken und damit die Angriffsfläche zu mindern.

Abgesehen von neuen KI-spezifischen Bedrohungen lässt sich KI im Unternehmenskontext allerdings auch zur Erhöhung des Sicherheitsniveaus einsetzen und kann damit zu einer wichtigen Stütze der IT-Sicherheit des Unternehmens werden. Die Fähigkeiten von KI zur Mustererkennung können dabei gewinnbringend zur Detektion von Cyberangriffen, zum Schutz der Endgeräte, oder auch zur Erkennung des ungewünschten Abflusses von Daten eingesetzt werden.

Insgesamt lässt sich damit festhalten, dass das Potenzial von KI im Unternehmenskontext sehr groß ist und der Nichteinsatz dieser Methoden auch wirtschaftliche Folgen haben kann. Spezifische Bedrohungen sind zu berücksichtigen und den daraus resultierenden Risiken mit den in diesem Dokument dargestellten Maßnahmen zu begegnen.

Eine erfolgreiche Umsetzung solcher Maßnahmen ist auch Voraussetzung für eine Zertifizierung von KI-Systemen, welche entscheidend ist, um deren Sicherheit zu validieren

und Vertrauen zu schaffen. Zahlreiche Zertifizierungsverfahren bewerten die Einhaltung von Sicherheitsstandards und Datenschutz, fördern Transparenz und gewährleisten die Robustheit gegen Cyberangriffe. Solche Zertifizierungsprozesse sind maßgebend für die Vertrauenswürdigkeit von KI, indem sie sicherstellen, dass Systeme ethischen Grundsätzen und Sicherheitsvorschriften entsprechen. Kurzum, Zertifizierungen dienen als Gütesiegel, welche die sichere und verantwortungsvolle Nutzung von KI erleichtern und unterstützen.

Bitkom vertritt mehr als 2.200 Mitgliedsunternehmen aus der digitalen Wirtschaft. Sie generieren in Deutschland gut 200 Milliarden Euro Umsatz mit digitalen Technologien und Lösungen und beschäftigen mehr als 2 Millionen Menschen. Zu den Mitgliedern zählen mehr als 1.000 Mittelständler, über 500 Startups und nahezu alle Global Player. Sie bieten Software, IT-Services, Telekommunikations- oder Internetdienste an, stellen Geräte und Bauteile her, sind im Bereich der digitalen Medien tätig, kreieren Content, bieten Plattformen an oder sind in anderer Weise Teil der digitalen Wirtschaft. 82 Prozent der im Bitkom engagierten Unternehmen haben ihren Hauptsitz in Deutschland, weitere 8 Prozent kommen aus dem restlichen Europa und 7 Prozent aus den USA. 3 Prozent stammen aus anderen Regionen der Welt. Bitkom fördert und treibt die digitale Transformation der deutschen Wirtschaft und setzt sich für eine breite gesellschaftliche Teilhabe an den digitalen Entwicklungen ein. Ziel ist es, Deutschland zu einem leistungsfähigen und souveränen Digitalstandort zu machen.

Herausgeber
Bitkom e.V.
Albrechtstr. 10 | 10117 Berlin

Ansprechpartner/in
Janis Hecker | Referent Künstliche Intelligenz
T +49 30 27576-239 | j.hecker@bitkom.org
Felix Kuhlenkamp | Bereichsleiter Sicherheitspolitik
T +49 151 18882-727 | f.kuhlenkamp@bitkom.org

Verantwortliches Bitkom-Gremium
AK Artificial Intelligence & AK Informationssicherheit

Autorinnen und Autoren
Dr. Frank Beer, Principal Artificial Intelligence, infodas GmbH (Kapitel 3) | Sebastian Gerlach, Senior Director Policy EMEA, Palo Alto Networks GmbH (Kapitel 3) | Dr. Daniel Gille, Referatsleiter Künstliche Intelligenz, Agentur für Innovation in der Cybersicherheit GmbH (Kapitel 2.1 und 2.4) | Jan Ibisch, Referat T 25 – Sicherheit in der Künstlichen Intelligenz, Bundesamt für Sicherheit in der Informationstechnik (Kapitel 2.3) | Syrko Kulas, Referent Schlüsseltechnologien, Agentur für Innovation in der Cybersicherheit GmbH (Kapitel 2.1 und 2.4) | Volker Reers, CEO & Founder, Qseidon GmbH (Kapitel 2.1 und 3) | Annegrit Seyerlein-Klug, intcube GmbH; Dozentin für IT-Security und Security Management, Technische Hochschule Brandenburg (Kapitel 2.2 und 2.5) | Kai Pascal Beerlink, Junior Research Fellow, Brandenburgisches Institut für Gesellschaft und Sicherheit (BIGS) (Projektkoor-dination und Einzelbeiträge)

www.bitkom.org