Hochschule
Albstadt-Sigmaringen
Albstadt-Sigmaringen University

BOSCH

Hochschule Albstadt-Sigmaringen

Fakultät Informatik

Studiengang IT-Security

Bachelor-Thesis

Modell-Diebstahl für Zeitreihenprognosen in Bezug auf ein Cybersecurity Governance Framework für Künstliche Intelligenz

Betreuer:	**Prof. Dr. Nils Herda**
Zweit-Betreuer:	**Dr. Jesus Luna Garcia,** **Cloud und KI-Security Experte,** **Robert Bosch GmbH**
Vorgelegt von:	**Valentin Acker**
Matrikelnummer:	**89175**

Ort:	**Mössingen**
Abgabetermin:	**23.08.2022**

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie die aus fremden Quellen direkt oder indirekt übernommenen Stellen/Gedanken als solche kenntlich gemacht habe. Diese Arbeit wurde noch keiner anderen Prüfungskommission in dieser oder einer ähnlichen Form vorgelegt. Sie wurde bisher auch nicht veröffentlicht.

Hiermit stimme ich zu, dass die vorliegende Arbeit von der Prüferin/ dem Prüfer in elektronischer Form mit entsprechender Software auf Plagiate überprüft wird.

Mössingen, 22.08.2022
Ort, Datum

Valentin Acher
Unterschrift des Studierenden

Dieses Projekt wurde mit Mitteln aus dem Forschungs- und Innovationsprogramm Horizont 2020 der Europäischen Union unter der Finanzhilfevereinbarung Nr. 952633 gefördert.

Abstract

Künstliche Intelligenz intergiert sich immer weiter in unser Leben und bietet viele Chancen. Daher ist es notwendig zu verstehen, welche Gefahren von KI-Systemen ausgehen können, die mangelnde Sicherheit aufweisen. Die vorliegende Arbeit zielt darauf ab, die Gefahren und das Potential von Cyber-Sicherheit in diesem Themenbereich aufzuzeigen. Dabei wird auf die herkömmliche IT-Governance eingegangen und durch den Bezug auf die Künstliche Intelligenz erweitert. Durch die unterschiedlichen Perspektiven der Künstlichen Intelligenz für Cyber-Sicherheit, steigt der Bedarf an Governance für KI-Cyber-Sicherheit und wird deswegen thematisiert. Um die Notwendigkeit der Cyber-Sicherheit für Künstliche Intelligenz zu verdeutlichen, wurde ein Angriff auf mehrere KI-Modelle durchgeführt, die auf Zeitreihenprognosen spezialisiert sind. Diese sind in verschiedensten Gebieten von großer Bedeutung, da sie dabei helfen, die Zukunft besser abschätzen zu können und somit frühzeitig reagieren zu können. Das Ziel war es dabei, die Funktionsweise des KI-Modells auf ein Angreifer Modell zu übertragen. Diese Vorgehensweise ist nicht auf Zeitreihenprognosen spezifiziert, sondern kann auf unterschiedlichste KI-Modelle angewendet werden. Der durchgeführte Angriff zeigt, dass innerhalb von kurzer Zeit, durch erhaltene Prognosen aus einem originalen Modell, ein Angreifer Modell trainiert werden konnte. Weiterführende Forschung im Bereich der Angriffe auf Zeitreihenprognosen durch Künstliche Intelligenz könnte auf Invasion Angriffe oder auf Data Poisoning Angriffe ausgerichtet sein. So ist es möglich, das Bewusstsein für Angriffe auf Zeitreihenprognosen der Künstlichen Intelligenz weiter zu schärfen und die potenziellen Gefahren zu minimieren oder auch neue Ansätze für den Schutz von KI-Systemen zu finden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIC	Akaike-Informationskriterium
ARIMA	Autoregressive Integrated Moving Average
BSI	Bundesamt für Sicherheit in der Informationstechnik
CNN	Convolutional Neural Network
GAN	Generative Adversarial Networks
IT	Informationstechnologie
IDS	Intrusion Detection System
KI	Künstliche Intelligenz
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RNN	Recurrent Neural Network
XAI	Explainable Artificial Intelligence

1 Einführung

1.1 Motivation

Künstliche Intelligenz (KI) hat sich zur heutigen Zeit in viele Anwendungsgebiete etablieren können und unterstützt uns Menschen beinahe täglich. Es werden immer mehr Fortschritte erzielt und so entstehen immer weitere Möglichkeiten, neue KI-gestützte Produkte und Dienstleistungen zu entwickeln, die sich in unser Leben integrieren. Künstliche Intelligenz ist zu einem wichtigen Bestandteil in unserem Leben geworden. Vor allem Zeitreihenprognosen in Kombination mit Künstlicher Intelligenz können einen enormen Mehrwert liefern. Selbst in der Corona Pandemie konnte Künstliche Intelligenz verwendet werden, indem durch Zeitreihenprognosen die aktiven Infektionen vorhergesagt werden konnten. Dies konnte dazu beitragen, die Pandemie besser einschätzen zu können und den Verlauf sowie die Auswirkungen zu kontrollieren. Dennoch ist sie noch nicht an ihre Grenzen gestoßen und wird eine immer wichtiger werdende Rolle sowohl in der Wirtschaft als auch in der Gesellschaft einnehmen. (Vgl. Kumar & Susan, 2020: S. 2 f.)

Mit der Künstlichen Intelligenz wird versucht, den Fortschritt immer weiter voranzutreiben und durch die daraus resultierende Automatisierung immer effizienter zu werden. Jedoch wird durch die automatisierten Prozesse für den Großteil der Bevölkerung unklar, wie sich die Entscheidungen der Künstlichen Intelligenz zusammensetzen. Beispielsweise konnte bei einer Künstlichen Intelligenz, die Bilder von Tieren dem jeweiligen Tier zuordnet, durch kleine Veränderungen so abgeändert werden, dass eine falsche Klassifikation des zu erkennenden Tieres durchgeführt wurde. Diese Veränderungen sind so minimal, dass sie dem Auge des Menschen nicht auffallen. Somit kann nicht ohne weiteres nachvollzogen werden, warum welche Aktion von der Künstlichen Intelligenz ausgeführt wurde. Deshalb muss an der Erklärbarkeit von den Modellen und der Künstlichen Intelligenz an sich gearbeitet werden. (Vgl. Zhang & Li, 2020: S. 2579 f.)

Datenschutz wird immer wichtiger für die Menschen, jedoch ist durch die nicht ausreichende Sicherheit in der Künstlichen Intelligenz, der Datenschutz nicht gegeben. Zu der Nachvollziehbarkeit gehört auch die Beschaffenheit der Daten. Ein wichtiger Faktor hierbei ist, ob diese Daten ohne Befangenheit sind oder wie mit diesen Umgegangen wird. Ansonsten behandelt die Künstliche Intelligenz verschiedene Gruppen auf unterschiedliche Art und Weise. Dies kann in Vor- und Nachteilen diesen Gruppen gegenüber enden. Somit ist die Fairness von Künstlicher Intelligenz nicht gegeben. (Vgl. Executive Office of the President, 2016: S. 27 f.)

Aus technischer Perspektive spielt die Genauigkeit eine wichtige Rolle, da je nach Einsatzgebiet schwere Schäden und Kosten durch mangelnde Genauigkeit entstehen können. Deshalb sollte auf die Verlässlichkeit der Künstlichen Intelligenz geachtet werden, um Vorfälle zu minimieren. Darüber hinaus sollte ein solches System robust sein, um die Anzahl der Ausfälle so gering wie möglich zu halten. (Vgl. Executive Office of the President, 2016: S. 9)

Für die herkömmliche Technik haben sich über die Jahre verschiedene Methoden entwickelt, um diese in Bezug auf Cyber-Sicherheit vor potenziellen Angreifern als auch falscher Nutzung zu schützen. Die Gewährleistung von Verfügbarkeit, Integrität und Vertrauenswürdigkeit ist sowohl in der herkömmlichen Technik als auch in der Künstlichen Intelligenz essenziell. Jedoch sind die bereits entwickelten Methoden auch für die Künstliche Intelligenz weitestgehend geeignet. In diesem Bereich müssen weitergehende Schutzmechanismen entwickelt und angewendet werden, da sich neue und unbekannte Schwachstellen aufgetan haben, die böswillig ausgenutzt werden können. Deshalb müssen für den Schutz von KI-Systemen weitere Gegenmaßnahmen implementiert werden, um diesen zu verbessern. (Vgl. Sowa, 2020: S. 134 f.)

Alle diese Punkte helfen bei richtiger Umsetzung dabei, die Vertrauenswürdigkeit von Künstlicher Intelligenz zu steigern und sollte bei der weiteren Entwicklung der Künstlichen Intelligenz eine primäre Rolle spielen, vergleiche Abbildung 1: Schutz für Künstliche Intelligenz. Deshalb zielt diese Arbeit darauf ab, dass alle diese Ansätze ihre Daseinsberechtigung haben und dazu beitragen werden, Künstliche Intelligenz weiter voranzubringen und die Vertrauenswürdigkeit in diese zu steigern.

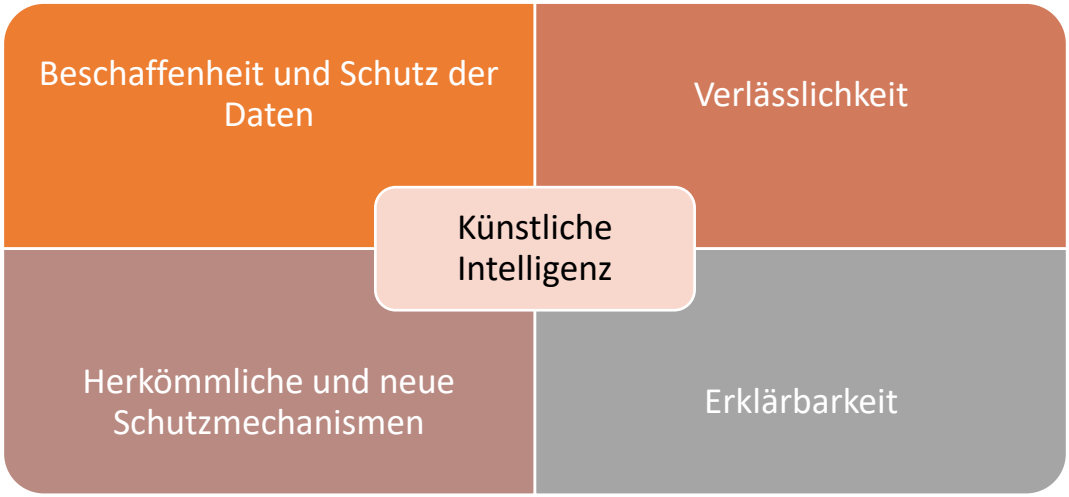

Abbildung 1: Schutz für Künstliche Intelligenz (Quelle: Eigene Darstellung)

1.2 Zielstellung

Das Ziel dieser Arbeit ist es, die verschiedenen Aspekte für ein KI-Cyber-Sicherheits-Governance-Framework zu untersuchen und zu analysieren. Einbezogen werden hierbei die Aspekte der Bedrohung, das Risiko als auch die Schutzvorkehrungen mit Bezug auf Künstliche Intelligenz. Des Weiteren werden Bewertungen der Künstlichen Intelligenz spezifischen Schwachstellen von Modellen des Machine Learnings durchgeführt, die Prognosen auf Grundlage von Zeitreihen erstellen. Dabei wird der Grad der Gefährdung solcher Modelle bewertet, um bei der Entwicklung neuer Methoden zum Schutz vor Angriffen zu unterstützen. Darüber hinaus wird ein experimentelles Benchmarking mit ausgewählten bestehenden und neuen Lösungen für diese spezielle Aufgabe durchgeführt. Schließlich werden die erlangten Erfahrungen und Ergebnisse in das Risikomanagement-/Governance-Framework für KI-Cyber-Sicherheit integriert.

1.3 Struktureller Aufbau

Diese Arbeit untergliedert sich in acht Kapitel. Das erste Kapitel gibt einen Überblick über die Motivation, die Ziele sowie den Aufbau der Arbeit. Im zweiten Kapitel wird in die Thematik der Künstlichen Intelligenz eingeführt. Näher erläutert werden hierbei die Entwicklung, die Arten, die Funktionalitäten, sowie die Anwendungsgebiete der Künstlichen Intelligenz. Im nachfolgenden Kapitel wird das Thema IT-Governance näher betrachtet und mit Cyber-Sicherheit verknüpft. Im vierten Kapitel wird auf die Perspektiven der Künstlichen Intelligenz mit Bezug auf Cyber-Sicherheit eingegangen. Die enthaltenen Kategorien dieses Kapitels werden untergliedert in Adversarial Künstliche Intelligenz, in die Künstliche Intelligenz für Cyber-Sicherheit und abschließend in Cyber-Sicherheit für Künstliche Intelligenz.

Im fünften Kapitel sollen die Aspekte der Cyber-Sicherheit-Governance für Künstliche Intelligenz analysiert und deren Bedeutung erkenntlich gemacht werden. Darunter fallen die Aspekte der Bedrohungsmodellierung, Risk Management Framework und die Schutzvorkehrungen für KI-Systeme. Danach folgt die Bedrohung des Modell-Diebstahls durch Zeitreihenprognose. In diesem Kapitel werden die Punkte Zeitreihenprognosen, Modell-Diebstahl und die Gegenmaßnahme AIShield an einem eigenen Modell abgehandelt.

Um für die Zukunft Künstliche Intelligenz abzusichern, werden im vorletzten Kapitel, Zukunft der Cyber-Sicherheit für Künstliche Intelligenz, die Aspekte Überwachung, Zertifizierung, Kennzeichnung und die Reaktion auf KI-Vorfälle dargestellt. Abschließend wird die Arbeit zusammengefasst und die daraus resultierenden Ergebnisse festgehalten. Aus den daraus entstandenen Endergebnissen wird ein Ausblick des Themas für die Zukunft gegeben.

2 Künstliche Intelligenz

2.1 Kapitelübersicht

Künstliche Intelligenz ist keine neue Erfindung des 21. Jahrhunderts, dennoch ist sie nun nicht mehr wegzudenken. Sie findet in unterschiedlichsten Anwendungsgebieten Einsatz und das Potential ist noch nicht ausgeschöpft. Sie zeichnet sich dadurch aus, dass es unterschiedlichste Ansätze gibt, die für die Anwendungsgebiete geeignet sind. Dabei kann Künstliche Intelligenz bei Tätigkeiten unterstützen, diese selbst ausüben oder in Zukunft eventuell den Menschen selbst übertreffen.

2.2 Entwicklungshistorie

Künstliche Intelligenz ist kein neues Phänomen, die Idee dahinter existiert seit über einem halben Jahrhundert. Das erste Mal als von Künstlicher Intelligenz gesprochen wurde, war im Jahr 1956. In diesem Jahr fand ein Forschungsprojekt statt, dass die Möglichkeiten von Künstlicher Intelligenz thematisierte. Als Basis der Studie diente der Gedanke, dass alle Möglichkeiten des Lernens oder auch der Intelligenz so dargestellt werden können, dass eine Maschine diese verwenden kann. Dieser Punkt stellt den Beginn der Forschung für Künstlicher Intelligenz dar. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 32 ff.)

Sechs Jahre zuvor gab Alan Turing in seiner Arbeit eine Anregung für das unbekannte Themengebiet, in dem er die folgende Frage stellte, „Können Maschinen Denken?“ (TURING, 1950: S. 433). Er kam früh auf den Entschluss, dass die Antwort auf diese Frage nicht möglich wäre, da Menschen und Maschinen auf unterschiedlicher Art und Weise Denken. Die menschliche Definition für „denken“ wäre somit nichtzutreffend und die Frage muss auf anderem Wege angegangen werden. Dafür entwickelte Turing das Imitation Game, damit versuchte er eine Untersuchung durchzuführen, die zeigen sollte, ob sich eine Person täuschen lässt, indem eine Maschine das Denken einer Person nachahmt. Eine Maschine gelte nur dann als intelligent, wenn die Antworten eines Computers mit denen einer Testperson übereinstimmen und somit nicht zu unterscheiden sind. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 33; TURING, 1950: S. 433 f.)

Nach dem der Antrag 1956 gestellt wurde, erlebte die Künstliche Intelligenz einen steilen Aufschwung. Es wurden viele KI-Programme entwickelt und dazu wurden weitere Untersuchungsergebnisse veröffentlicht. Damit dieser Aufschwung stattfinden konnte, wurde viel in das neue Themengebiet investiert. Jedoch hielt dieser Aufschwung nur bis in die 1970er-Jahre an. Die Entwicklung konnte nicht den ausreichenden Fortschritt erzielen und somit

wurden die finanziellen Investitionen gekürzt. Auch das öffentliche und wirtschaftliche Interesse flachte ab und daraus entstand der KI-Winter. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 33 f.)

In diesem Zeitraum breitete sich Enttäuschung in Bezug der Künstlichen Intelligenz aus. Begeisterung für das Thema kam erst im Jahr 1997 erneut auf, als der Schachcomputer Deep Blue gegen den Schachweltmeister gewinnen konnte. Seit diesem Zeitpunkt ist das Interesse an der Künstlichen Intelligenz rasant angestiegen und wurde immer populärer, bis im 21. Jahrhundert eine regelrechte Begeisterung um das Thema entstanden ist. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 33-35)

2.3 Systematisierung

Künstliche Intelligenz ist kein Ansatz für ausschließlich einen Bereich, vielmehr bietet sie durch verschiedene Methoden, Möglichkeiten um in unterschiedlichste Anwendungsgebiete eingegliedert zu werden. Jedoch variieren die verwendeten Methoden in ihrer Leistungsstärke, so kann die Künstliche Intelligenz in zwei beziehungsweise drei Teilbereiche untergliedert werden. Zu diesen zählen die schwache und die starke Künstliche Intelligenz. Die dritte und höchste Form ist die Artificial Superintelligence, jedoch ist bei dieser Form nicht sicher, wann und ob es sie jemals geben wird. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 37; Fang et al., 2018: S 5 f.)

Die schwache Künstliche Intelligenz suggeriert ihr Handeln so, als wäre sie intelligent. Dabei genügt es, wenn sie bestimmte Fälle abarbeiten und Menschen bei bestimmten Tätigkeiten unterstützen kann. So entsteht der Eindruck der Intelligenz, obwohl kaum Merkmale vorhanden sind, die mit menschlicher Intelligenz gleichgesetzt werden können. Damit soll nicht signalisiert werden, dass schwache Künstliche Intelligenz tatsächlich schwach ist, viel mehr soll hervorgehoben werden, dass sie sich auf die Spezialisierung konzentriert. Sie nutzt ihre Fähigkeiten, um für die Anwendungsbereiche, für die, die Applikationen entwickelt wurden, zu schärfen und sich für diese zu spezialisieren. Zur heutigen Zeit fallen die meisten Anwendungen der Künstlichen Intelligenz unter den Teilbereich der „schwachen“ Künstlichen Intelligenz. (Vgl. Fang et al., 2018: S. 5 f.)

Starke Künstliche Intelligenz oder auch als Artificial General Intelligence bezeichnet, soll dem Menschen ebenbürtig sein. Sie soll dasselbe Intelligenz-Niveau erreichen, wie die menschliche Intelligenz. Um dies umsetzen zu können, muss die Maschine logisch denken und lernen können. Jedoch ist es für uns noch nicht möglich, diese Fähigkeiten richtig umzusetzen, denn kein aktueller Algorithmus ist dazu in der Lage, Artificial General Intelligence abzubilden. Es

gibt Ansätze zu Theorien, die jedoch selbst für Supercomputer nicht handhabbar sind. Dennoch können diese Ansätze für die Zukunft hilfreich sein, indem sie für die Entwicklung der Algorithmen hinzugezogen werden können. (Fang et al., 2018: S. 5f.; L. N. Long & C. F. Cotner, 2019: S. 1 f.)

Um starke Künstliche Intelligenz zu überprüfen, wird häufig der Turing Test verwendet vergleiche Unterkapitel 2.2 Entwicklungshistorie. Jedoch wird dieser nicht von allen Menschen unterstützt, da die Meinungen dies bezüglich zwiegespalten sind. (Vgl. Saygin et al., 2003: S. 464 f.)

Die nächste Stufe der Künstlichen Intelligenz ist die Artificial Superintelligence, diese stellt zum aktuellen Zeitpunkt ausschließlich eine Hypothese dar. Ziel ist es alle kognitiven Fähigkeiten des Menschen zu übertreffen. Deswegen können keine genauen Prognosen abgegeben werden, wann und ob eine solche Superintelligenz in der Zukunft erreicht werden kann. Um diese Stufe zu erreichen, muss zunächst die vorherige Stufe der Artificial General Intelligence erreicht werden. So kann sich diese weiterentwickeln und unter den richtigen Umständen die höchste Ebene erreichen. (Vgl. Bostrom: S. 22; Fang et al., 2018:S. 6)

2.4 Leistungsbreite

Künstliche Intelligenz bietet für unterschiedliche Komplexitätsgrade eine Vielzahl an Verfahren an, um diese zu bewältigen. Der Lösungsweg für die entstehenden Resultate gilt meist als irrelevant, da Systeme der Künstlichen Intelligenz anhand ihrer Leistung und dem daraus resultierenden Ergebnis gemessen werden. Ein Faktor, der zu dieser Herangehensweise beiträgt ist, dass der interne Ablauf eines KI-Systems nicht nachvollziehbar sein kann und deshalb schwer als Bewertung hinzugezogen werden kann. (Deutscher Dialogmarketing Verband e. V., 2019: S. 38)

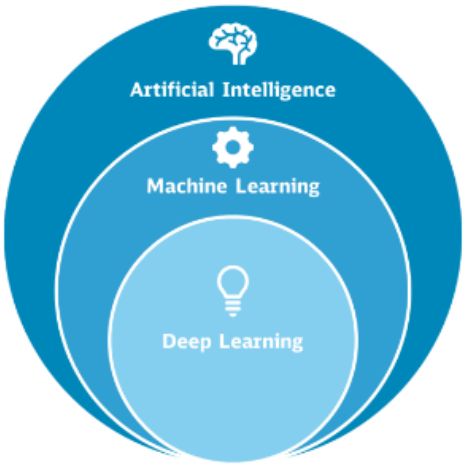

Abbildung 2: Funktionalitäten der Künstlichen Intelligenz (Quelle: Bauer, 2021: S. 17)

Im Folgenden wird auf Machine Learning, Deep Learning, sowie Cognitive Computing eingegangen, die, die Funktionalitäten der Künstlichen Intelligenz darstellen. Diese stehen in Beziehung zueinander, vergleiche Abbildung 2: Funktionalitäten der Künstlichen Intelligenz auf Seite 6. Bei diesen Bereichen ist zu beachten, dass sie sich in großem Maß in der Klarheit des Anwendungszwecks sowie dem Grad der Autonomie unterscheiden. Zweck und Autonomie stehen somit in Beziehung und haben gegenseitige Auswirkung. Eine Funktionalität deren Zweck festgelegt ist, besitzt einen kleinen Anteil an Autonomie. Dies ist auch vice versa festzustellen. Auch die Innovation ist in diesem Aspekt der Künstlichen Intelligenz integriert, je innovativer eine Funktionalität ist, umso größer ist der Grad der Autonomie bei der Anwendung. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 38)

Machine Learning

Machine Learning stellt einen der wichtigsten Ansätze für Künstliche Intelligenz dar. Außerdem dient dieser Ansatz als Treiber der aktuellen Fortschritte, sowie der kommerziellen Verwendung der Künstlichen Intelligenz. In diesem Teilaspekt der Künstlichen Intelligenz wird versucht, neues Wissen zu generieren, dass auf vorhandenem Wissen basiert. Hierbei wird ein Datensatz verwendet, der als Grundlage dient. Daraus soll versucht werden, eine Regel oder Verfahren abzuleiten, mit dem die verwendeten Daten erklärt oder Zukunftsprognosen abgegeben werden können. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 39; Executive Office of the President, 2016: S. 8)

Ein Vorteil, der aus Machine Learning hervorgeht ist, dass selbst bei Situationen in denen es schwer ist Regeln zu finden, die dabei helfen sollen ein Problem zu lösen, durch die Verwendung von Machine Learning gefunden und angewendet werden können. Durch den dabei verwendeten statistischen Prozess wird versucht, mit den bereitgestellten Daten ein künstliches System zu trainieren, um somit vorkommende Muster oder besondere Merkmale zu erkennen. Durch diese sollen unbekannte Daten analysiert werden, um die Lösung für die jeweiligen Probleme zu finden. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019: S. 39; Executive Office of the President, 2016: S. 8 f.)

Für die Anwendung von Machine Learning wird ein Datensatz verwendet, der vergangene Werte enthält. Dieser Datensatz wird aufgeteilt, in einen Trainingsdatensatz und in einen Testdatensatz, der letztere dient zu Validierung der entstehenden Ergebnisse. Der Anwender wählt ein passendes Modell, das Regeln mit anpassbaren Parametern darstellt. Auch wird eine Zielfunktion erstellt, anhand dieser die Parameter angepasst werden können, um ein genaueres Ergebnis zu erhalten. Typisch für das Training des Modells ist ein Belohnungssystem. Dieses

verteil Belohnungen bei dem Verwenden von einfacheren Regeln und für eine hohe Übereinstimmung mit dem Trainingsdatensatz. Während dem Trainings-Prozess werden die Parameter so angepasst, dass die zuvor erstellte Zielfunktion maximiert wird. Jedoch stellt dieser Schritt auch den schwierigsten im ganzen Prozess dar. (Vgl. Executive Office of the President, 2016: S. 9)

Um die Genauigkeit des erstellten Modells zu überprüfen, wird der Testdatensatz als Basis verwendet. Wenn die Genauigkeit dem gewünschten Wert entspricht, können zukünftige Daten verwendet werden, die für das Modell fremd sind. Das Modell soll die neuen Daten mit der gleichen Genauigkeit verwerten können, wie die vorherigen Daten. (Vgl. Executive Office of the President, 2016: S. 9)

Das Ziel von Machine Learning ist es, ein Modell zu trainieren, dass nicht nur für Trainingsdaten eine hohe Genauigkeit aufweisen soll, sondern auch für zukünftige Daten, die dem Modell unbekannt sind. So kann Machine Learning nicht als Lösungsalgorithmus für ein spezifisches Problem betrachtet werden. Vielmehr stellt es einen allgemeinen Ansatz für unterschiedlichste Problemstellungen dar, für die Daten vorhanden sein müssen. Die nachfolgende Abbildung 3 zeigt den Zyklus, der bei der Implementierung und Verwendung eines KI-Modells verwendet wird. (Vgl. Executive Office of the President, 2016: S. 9)

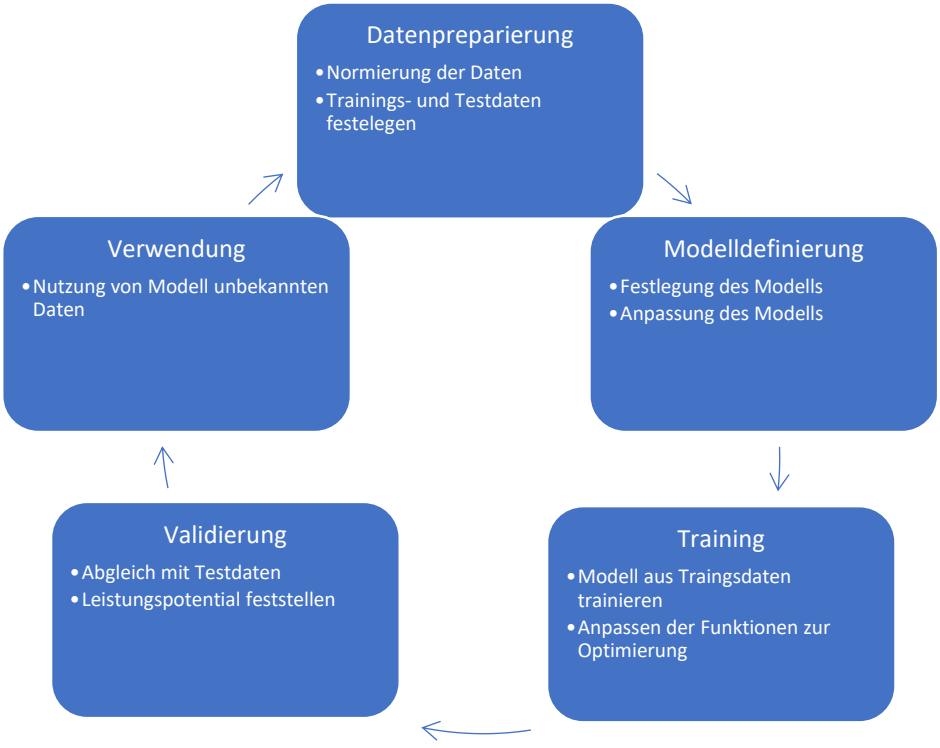

Abbildung 3: Implementierung KI-Modell-Zyklus (Quelle: Eigene Darstellung)

Deep Network Learning

Deep Network Learning bzw. auch als Deep Learning bezeichnet, stellt eine Sonderform von Machine Learning dar. Die Arbeitsweise dieser Form ist dem menschlichen Gehirn nachempfunden. Dabei sollen Einheiten so strukturiert werden, sodass diese wie die Neuronen des Gehirns agieren. So soll für jedes „Neuron“ der Anwendung eine Eingabe stattfinden, die einen Ausgabewert erzeugt und diesen an das nächste Neuron weiterleitet. Das imitierte neuronale Netz kann wenig bis zu großen Mengen an Neuronen enthalten, die in unterschiedlich ausgeprägten Schichten vertikal und horizontal angeordnet werden können. So sollen durch die Vielzahl der Schichten und den enthaltenen Einheiten, die großen Datenmengen analysiert und draus extrem komplexe und präzise Muster erkannt werden. (Vgl. Executive Office of the President, 2016: S.9 f.)

Durch den technischen Fortschritt konnten die schnelleren Computersysteme mit neuen Theorien des Aufbaus und Training dieser Methode miteinander verknüpft werden. Daraus resultierten größere Deep Learning Networks, die einen großen Beitrag zu der Entwicklung von Künstlicher Intelligenz leisten konnten. Dieser Beitrag konnte nicht nur die Entwicklung voranbringen, sondern auch mehr Menschen für Machine Learning begeistern. (Vgl. Executive Office of the President, 2016: S. 9 f.)

Cognitive Computing

Cognitive Computing stellt den Ansatz dar, der verwendet wird, um intelligente Systeme zu befähigen, damit sie mit Menschen und anderen intelligenten Systemen auf natürliche Weise interagieren können. So sollen diese Systeme Aufgaben und Entscheidungen übernehmen, die zum heutigen Zeitpunkt weitestgehend von Menschen ausgeübt und getroffen werden. Ein markantes Merkmal von Cogntitive Computing ist, dass die Systeme menschliche Eigenschaften verwenden, um mit den anfallenden Problemen umgehen zu können. (Vgl. Deutscher Dialogmarketing Verband e. V., 2019; Haluk Demirkan et al., 2017: S. 16 f.)

Die Besonderheit bei dieser Methode ist, dass die Systeme nicht auf verschiedene Szenarien programmiert werden. Sie lernen aus den Interaktionen mit Menschen oder anderen intelligenten Systemen und können aus diesen Erfahrungen, Folgerungen ableiten. So machen sich die Systeme ihre Umgebung zu Nutze und erzielen somit ein großes Maß an Autonomie, die einen großen Schritt in die Künstliche Intelligenz darstellt. (Vgl. Haluk Demirkan et al., 2017: S. 16)

Diese Dienstleistungen haben schon Einzug in das Leben vieler Menschen genommen. Trotzdem kann beim aktuellen Stand nicht von einer hochentwickelten Künstlichen Intelligenz gesprochen werden, dennoch kann sie einen großen Teil dazu beitragen, Kosten zu reduzieren, sowie die Effizienz zu steigern. Deshalb zielt Cognitive Computing darauf ab, die Produktivität als auch die Kreativität von Organisationen und Individuen zu steigern. Für die Zukunft können durch den verfolgten Ansatz neue Themenbereiche und Gebiete erschlossen werden, die zu heutigen Zeit noch nicht bekannt sind. (Vgl. Haluk Demirkan et al., 2017: S. 17)

2.5 Anwendungsgebiete

Viele Menschen hatten schon Kontakt mit Künstlicher Intelligenz, obwohl sie dies eventuell nicht aktiv mitbekommen haben. Künstliche Intelligenz hat in vielen Bereich Einzug in unser Leben gefunden, denn sie kann in fast allen Lebenssituationen unterstützen und ist in den meisten Arbeitsbereichen einsetzbar. Somit erobert sie jegliche Wirtschaftsbranchen. (Vgl. Wirtz & Weyerer, 2019: S. 1 f.)

Mit einer Form der Künstlichen Intelligenz sind die meisten Menschen bereits in Kontakt gekommen oder haben davon gehört. Dies sind KI-basierte Sprachassistenten zu denen Anwendungen gehören wie, Google Assistant, Alexa (Amazon), Siri (Apple), selbst in Automobilen wird diese Technik verwendet (Vgl. Wirtz & Weyerer, 2019: S. 5). Diese gehören zu der Kategorie der Smarten Geräte, Anlagen und Umgebungen. Diese Kategorie wird häufig mit Prozessüberwachung, der Steuerung und der Unterstützung des Managementbereichs in Verbindung gebracht. In einem nicht-industriellen Umfeld werden die oben genannten Sprachassistenten für die vereinfachte Nutzung von smarten Geräten verwendet. Darüber hinaus sind im privaten Bereich weitere smarte Geräte anzutreffen, dazu gehören auch intelligente Gebäudesteuerung oder in diversen Wearables. (Vgl. Dirk Hecker et al., 2018: S. 32 f.)

Auch in einem größeren Ausmaß kann das öffentliche Leben von Künstliche Intelligenz unterstützt werden. Hiermit können intelligente Ampelschaltungen den Verkehrsalltag regulieren und somit ein vorausschauendes Verkehrsmanagement für eine intelligente Stadt gewährleisten. In dieser kann selbst das Stromnetz intelligent sein. Um die Möglichkeiten sowohl in der Industrie als auch im öffentlichen- bzw. privaten Raum erfolgreich implementieren zu können, müssen noch weitere Fortschritte in den Bereichen des Machine Learnings, der natürlichen Sprachverarbeitung als auch Bilderkennung erzielt werden. (Vgl. Dirk Hecker et al., 2018: S. 28 f.)

Während die verwendete Technik in Transportmittel immer intelligenter wird, so werden diese an sich intelligenter, indem diese automatisiert werden. So können Fahrassistenzsysteme während der Verwendung den Fahrzeugführer unterstützen und manche Aufgaben autonom übernehmen. Jedoch sollen für die Zukunft nicht nur einzelne Aufgaben autonom durchgeführt werden, sondern das Fahrzeug soll die Fahraufgabe im vollen Umfang übernehmen und somit den Fahrzeugführer ablösen. Dafür sollen die erlangten Echtzeitdaten von verschiedensten Sensoren entnommen und die davon betroffenen Systeme direkt angesprochen werden. Auf diese Weise würde nicht nur die Mobilität auf dem Land revolutioniert werden, sondern auch zu Wasser und in der Luft. (Vgl. Dirk Hecker et al., 2018: S. 24 f.)

In Industrieanlagen werden seit den 1970er Jahren Roboter verwendet, die in einem extra abgesicherten Bereich repetitive Arbeit verrichten. Mit dem technischen Fortschritt konnten diese Roboter immer weiter in das Umfeld von menschlichen Arbeitern integriert werden, ohne dass für diese Gefahr besteht. Die Roboter wurden in ihren Tätigkeiten sicherer, variabler in den Einsatzgebieten und selbständiger. Mit Hilfe der Künstlichen Intelligenz konnten den Robotern die autonome Komponente angeeignet werden. So können diese autonomen Roboter durch die Wahrnehmung ihrer Umgebung interaktiv auf verschiedene Vorfälle agieren und reagieren. Die unterstützenden Roboter können in die Bereiche Industrierobotik und Servicerobotik unterteilt werden. Im industriellen Bereich arbeiten die Roboter mit Menschen zusammen und steigern die Produktivität. Serviceroboter hingegen werden außerhalb der Industrie verwendet und sollen durch Dienstleistungen, das Wohlbefinden des Menschen verbessern. (Vgl. Dirk Hecker et al., 2018: S. 12-18)

Eine weitere große Kategorie, in der die Anwendungsbereiche der Künstlichen Intelligenz liegen, sind kognitive Assistenten. Diese dienen dazu, Menschen bei Aufgaben zu unterstützen oder diese Arbeit komplett zu übernehmen, die im kognitiven Bereich liegen. So sollen diese Assistenten in textueller oder sprachlicher Form interagieren und so bei einem Entscheidungsfindungsprozess helfen. Aktuell werden vermehrt in Bereichen der Konsumelektronik, Kundenservice, Medizin-, Bank-, Finanz-, Versicherungswesen etc. diese kognitiven Assistenten zum Einsatz gebracht. Durch diese Mithilfe wird Entlastung bei simplen Anfragen und Routineaufgaben geboten und Self-Service als auch Rund-um-die-Uhr-Erreichbarkeit ermöglicht. Die Verwendung von solchen Chatbots findet hohe Akzeptanz unter den Kunden, da einfache Fragen vorher geklärt werden können. (Vgl. Dirk Hecker et al., 2018: S. 36-40)

Viele dieser Gebiete bringen nicht nur Fortschritt in einem spezifischen Anwendungsbereich, die meisten überschneiden sich und entwickeln die anderen Bereiche simultan mit. Es gibt viele weitere Bereiche, die durch Künstliche Intelligenz profitieren können. Von der Landwirtschaft über das Gesundheitswesen bis zur Sicherheit kann alles einen Nutzen aus der Künstlichen Intelligenz ziehen. Dennoch sind die genannten Anwendungsgebiete nur die Spitze des Eisbergs und zukünftige Anwendungsgebiete können davon auch betroffen sein.

2.6 Zusammenfassung

Künstliche Intelligenz ist ein Phänomen, das schon über ein halbes Jahrhundert bekannt ist. Dennoch ist es ein sehr aktuelles Thema und integriert sich immer mehr in unser Leben. Die meisten Anwendungen, die Künstliche Intelligenz verwenden stehen weitestgehend am Anfang des Möglichen. Ob sich die Künstliche Intelligenz soweit entwickeln kann, dass sie den Menschen als Artificial Superintelligence übertreffen wird, kann bisher noch nicht prognostiziert werden. Jedoch lässt sich durch die breite Vielfalt, die sich in den unterschiedlichen KI-Modellen widerspiegelt erkennen, welches Potential davon ausgeht und in welchen Bereichen dies bereits Anwendung finden und noch finden werden.

3 IT-GRC und Cyber-Sicherheit

3.1 Kapitelübersicht

Durch den technologischen Fortschritt der letzten Jahrzehnte wurden uns viele neue Möglichkeiten und Wege ermöglicht. Neue Produkte und Dienste werden zum Wegbegleiter und somit zum Bestandteil des alltäglichen Lebens. Unsere Abhängigkeit vom technologischen Bereich wächst stetig und so auch die wichtige Bedeutung der Cyber-Sicherheit. Immer mehr personenbezogene Daten werden via Internet geteilt. Die daraus resultierende breitgefächerte Vernetzung lässt die Wahrscheinlichkeit Opfer von Cyberkriminalität oder Cyberangriffen zu werden, signifikant ansteigen und stellt ein immer größer werdendes Problem dar. (Vgl. European Court of Auditors, 2019: S. 4)

3.2 Einführung

Die Cyber-Sicherheit besitzt keine einheitliche Definition, die in der Allgemeinheit anerkannt ist. Im Themenpapier des Europäischen Rechnungshofes wird der Begriff wie folgt beschrieben, „grob gesprochen handelt es sich um alle Vorkehrungen und Maßnahmen zum Schutz von Informationssystemen und deren Nutzern vor unbefugten Zugriffen, vor Angriffen und vor Schaden, um die Vertraulichkeit, Integrität und Verfügbarkeit von Daten zu gewährleisten“ (European Court of Auditors, 2019: S. 7).

Neben der Erkennung und Vorbeugung von Cybervorfällen, gehören zu der Cyber-Sicherheit auch die aktiven Maßnahmen gegen diese Vorfälle sowie die Erholung und Ausbesserung. Grundsätzlich müssen solche Vorfälle keinen Vorsatz haben und können bereits durch Unachtsamkeit bei der Weitergabe von Daten entstehen. Vorsätzliche Angriffe können stark im Ziel und den dahinterliegenden Absichten variieren. Dabei können sowohl Privatpersonen, Unternehmen als auch kritische Infrastrukturen und demokratische Prozesse in das Visier von Cyberkriminellen gelangen. (Vgl. European Court of Auditors, 2019: S. 7)

Jedes neue Gerät, das mit dem Internet oder mit anderen Geräten in verbindungtreten kann, stellt eine potenzielle Angriffsmöglichkeit dar und vergrößert somit die Angriffsfläche für Angreifer. Der technische Fortschritt bringt die Digitalisierung in viele Bereiche und daraus konnten sich neue Möglichkeiten entwickeln. Dazu gehört das Internet der Dinge, die Cloud und Big Data. All diese Bereiche können ein exponentielles Wachstum vorweisen und somit wächst die potenzielle Angriffsfläche der Angreifer stetig mit. Deutlich zu sehen ist dies im Umfeld vom Internet der Dinge. Hier werden in großen Mengen Systeme in Umlauf gebracht, die wenig oder sogar keine Schutzmechanismen integriert haben. So wird nicht nur die

Angriffsfläche vergrößert, sondern auch der Angriff auf diese Systeme vereinfacht. (Vgl. European Court of Auditors, 2019: S. 9; Maik Morgenstern et al., 2021: S. 103 f.)

Die Angriffe von Cyberkriminellen werden immer raffinierter und komplexer, deshalb können die Vorfälle Auswirkungen auf globaler Ebene bewirken. Dabei nutzen die Cyberkriminellen Verschleierungstaktiken mit deren Hilfe diese sehr schwer ermittelt werden können und somit unentdeckt bleiben. Einzelpersonen, kriminelle Vereinigungen aber auch Hacktivisten oder Staaten können aus den unterschiedlichsten Motiven, Cyberangriffe verwenden, um ihre Ziele zu erreichen. (Vgl. European Court of Auditors, 2019: S. 7 f.)

Cyber-Sicherheit stellt einen Aspekt dar, der von vielen Unternehmen wenig oder keine Berücksichtigung findet. Nur 20% der europäischen Unternehmen konnten im Jahr 2016 keinen Cyber-Sicherheits-Vorfall verzeichnen. Alle andern wurden mindestens mit einem solchen Vorfall konfrontiert. Durch diese enorme Anzahl ist der durch die Angriffe resultierende wirtschaftliche Schaden zwischen 2013 und 2017 auf das Fünffache des Wertes angestiegen. Trotz dieser deutlichen Zahlen hat sich die Anschauungsweise der Cyber-Sicherheit als auch das Risikobewusstsein in den Unternehmen kaum verändert. Mehr als die Hälfte der Unternehmen sind sich der ausgehenden Gefahr durch Cyberkriminalität kaum oder nicht bewusst. 60% der Unternehmen haben keine Abschätzung für mögliche finanzielle Verluste durchgeführt. (Vgl. European Commission, 2017: S. 1; European Court of Auditors, 2019: S. 10)

Der finanzielle Schaden, der durch Cyberangriffe verursacht werden kann, steht kaum in Relation zu den Kosten, die für präventive Maßnahmen, Ermittlung als auch das Ausbessern der angefallenen Schäden entstehen. Ein Distributed-Denial-of-Service-Angriff kann durch viele Anfragen die Verfügbarkeit von Services enorm einschränken oder sogar für einen Absturz sorgen. Solch ein Angriff kann von Leuten mit wenig oder keinem technischen Wissen gestartet werden und das für einen Preis von 15 Euro pro Monat. (Vgl. European Court of Auditors, 2019: S. 10; Europol, 2018)

Durch die schwerwiegenden Cyberangriffe im Jahr 2017, die, die meisten Länder betroffen haben, wurde das Risikobewusstsein auf politischer Ebene verstärkt wahrgenommen. Dies brachte die Cyber-Sicherheit zurück in die Politik. Durch diesen Ansatz soll verstärkt ein Ansatz für die Cyber-Sicherheit implementiert werden. Da Cyber-Sicherheit ein weitreichendes und komplexes Thema darstellt, muss dies je nach Gebiet auf andere Weise integriert werden. Deshalb sind richtige Maßnahmen und die Verwendung der Cyber-Sicherheit essenziell. (Vgl. European Court of Auditors, 2019: S. 10)

3.3 Betriebliche Maßnahmen

Um einen Nutzen aus Cyber-Sicherheit ziehen zu können, muss darauf geachtet werden, dass je nach Anwendungsgebiet, die passenden Maßnahmen gewählt werden. Bei diesen muss darauf geachtet werden, dass sie richtig verwendet werden, um die Verfügbarkeit, Integrität und auch Vertraulichkeit gewährleisten zu können. Wegen den verschiedenen Bedrohungsarten sind die genannten Sicherheitsgrundsätze unterschiedlich stark den Gefahren ausgesetzt.

Abbildung 4: Bedrohungen und deren Auswirkung auf Sicherheitsgrundsätze (Quelle: European Court of Auditors, 2019: S. 8)

Abbildung 4 verdeutlicht die unterschiedlichen Auswirkungen je nach Bedrohungsart. Dabei stellt ein unbefugter Zugriff eine Gefahr für alle drei Sicherheitsgrundsätze dar, die durch das Ausrufezeichen dargestellt wird. Bei der Preisgabe von Daten, ist die Vertraulichkeit dieser nicht mehr zu gewährleisten, da sie von Dritten eingesehen werden können. Dennoch sind die Verfügbarkeit und Integrität bei dieser Bedrohungsart nicht beeinträchtigt, dies ist durch das Schloss-Symbol dargestellt. Bei der Datenveränderung hingegen steht die Integrität im Fokus, da die Daten verändert werden und durch diese Veränderung unbrauchbar werden oder falsche Informationen liefern. Bei der Datenzerstörung und Denial-of-Service wird die Verfügbarkeit gefährdet. Die Zerstörung dieser Daten kann zur Folge haben, dass Funktionen nicht zur Verfügung stehen. Bei der Denial-of-Service Methode wird versucht, die Verfügbarkeit von Systemen durch eine Vielzahl von Anfragen zu unterbinden. (Vgl. European Court of Auditors, 2019: S. 8)

Im weiteren Verlauf des Unterkapitels werden Maßnahmen genannt, wie verschiedene Bedrohungen durch die richtige Verwendung von Cyber-Sicherheit vorgebeugt und die Schäden verhindert werden können. Social Engineering ist ein wesentlicher Bestandteil von Cyber-Kriminalität, der kontinuierlich an Bedeutung gewinnt. Social Engineering gibt es in

verschiedensten Formen, die beliebteste davon stellt Phishing über E-Mail-Dienste dar. Mit dieser Methode wird versucht, unterschiedliche Ziele zu erreichen. Cyber-Kriminelle wollen damit personenbezogene Daten erlangen, Konten übernehmen, Identitäten von Personen übernehmen oder auch Geld transferieren. Dabei kann eine kleine Unachtsamkeit bei dem Lesen einer E-Mail schwerwiegende Folgen haben. (Vgl. Internet Organised Crime Threat Assessment (IOCTA) 2018 | Europol, 2022: S. 8 ff.)

Obwohl nur ein kleiner Anteil der Personen auf eine Phishing-Mail reagiert, ist ein Click auf einen Link ausreichend, um eine komplette Organisation zu kompromittieren. Um die daraus resultierenden Schäden eingrenzen oder komplett verhindern zu können, müssen technische als auch weiterbildende Maßnahmen getroffen werden. Eine technische Maßnahme, die sehr wirkungsvoll ist, ist das Verwenden von einem E-Mail-Filter. Durch diesen werden die Inhalte einer E-Mail überprüft und anschließenden weitergeleitet an den eigentlichen Empfänger, wenn kein verdacht vorliegt. Falls ein Verdacht vorliegt, wird die vermutlich bedrohliche E-Mail nicht an den Empfänger weitergeleitet. Jedoch entwickeln sich Phishing-E-Mails stetig weiter und erschweren somit das Detektieren von der ausgehenden Bedrohung dieser E-Mails. So ist es möglich, dass diese nicht erkannt und trotz enthaltener Gefahr, an die eigentlichen Empfänger weitergeleitet werden. (Vgl. Almomani et al., 2013: S. 2071 f.)

Dieses Problem lässt sich nicht durch eine technische Maßnahme lösen. Zu diesem Zeitpunkt ist der Empfänger auf sich gestellt und muss eine E-Mail richtig einschätzen können. Deshalb müssen Schulungen angeboten werden, die auf diese Gefahren aufmerksam machen. Diese sollten Exemplare enthalten, die auf unterschiedlichen Schwierigkeitsstufen basieren. Dadurch wird gewährleistet, dass sich die Teilnehmer nicht durch einfache Exemplare täuschen lassen und somit ein falsches Bewusstsein durch Überschätzung dafür entwickeln. Auch sollte darauf geachtet werden, dass diese Schulungen mehrmals im Jahr durchgeführt werden, da das Erlernte schnell in Vergessenheit gerät. In diesen Schulungen soll nicht nur das Wissen zum Erkennen von Phishing vermittelt werden, sondern auch das Melden von verdächtigen E-Mails sollte im Gedächtnis bleiben. Ein Grund hierfür ist, dass Experten diese gemeldeten E-Mails untersuchen und bei falschem Verdacht, eine Freigabe für diese erteilen können. Wenn jedoch eine gefährliche E-Mail gemeldet wird, können die Experten reagieren und diese E-Mail aus dem System entfernen und dadurch andere Empfänger schützen. Auch können durch die erkannten Phishing Versuche, die Filter erweitert werden, damit diese neue Phishing-Methoden abwehren können. Diese Vorgehensweise ist ein guter Schutz vor Phishing-Angriffen, garantiert aber keine vollständige Sicherheit. Denn die Schwachstelle Mensch, kann in unterschiedlichen Fällen unabsichtlich einen Fehler machen. (Vgl. Singh et al., 2019: S. 454 f.)

Häufig wird Phishing dazu verwendet, um Ransomware auf verschiedenste Geräte aufzuspielen. Wenn Ransomware auf ein Gerät gelangt, wird dieses mit allem darauf gespeicherten Inhalt verschlüsselt. Die betroffene Person ist nicht in der Lage, selbstständig das System zu entschlüsseln. Um die gespeicherten Daten wiederzuerlangen, muss Lösegeld in Kryptowährung bezahlt werden, da diese bei richtiger Verwendung nicht nachverfolgt werden kann. Bei diesem Angriff kann ein sehr hoher wirtschaftlicher Schaden entstehen. Je nach betroffenem Bereich in einem Unternehmen, können die Maschinen stillstehen und dadurch einen enormen Schaden anrichten. (Vgl. Savita Mohurle & Manisha Patil, 2017: S. 1 f.)

Mit Hilfe der Maßnahmen gegen Phishing, wird simultan das Risiko einer Infizierung durch Ransomware reduziert. Dennoch sollte bei dieser Bedrohung zusätzlich darauf geachtet werden, dass bei einer Infektion umgehend der Vorfall gemeldet wird, um das Ausbreiten zwischen Systemen zu unterbinden und den Schaden einzudämmen. Darüber hinaus muss daran gedacht werden, dass Ransomware auch über andere Wege in ein Unternehmen gelangen kann. Einer dieser Wege ist über einen USB-Stick, der von einem Angreifer oder von einem unwissenden Opfer verwendet wird. Um diese Möglichkeit einzuschränken, können USB-Ports deaktiviert oder nur unter bestimmten Berechtigungen aktiviert werden. Außerdem ist es sinnvoll, schützenswerte Gegenstände in eigene Zonen zu separieren. Ausschließlich für Personen mit den erforderlichen Berechtigungen, darf der Zutritt gewährt werden. (Vgl. Savita Mohurle & Manisha Patil, 2017: S. 1 f.)

Um das Risiko für Schwachstellen innerhalb verschiedenster Anwendungen und Systemen gewährleisten zu reduzieren, empfiehlt es sich, ein aktives Patch-Management zu betreiben. So können Schwachstellen aus älteren Versionen geschlossen werden, indem die neuste Version aufgespielt wird. Deshalb ist die Durchführung eines Patch-Vorgangs zu einem frühen Zeitpunkt erstrebenswert. (Vgl. Huseyin Cavusoglu et al., 2006: S. 1 f.)

Denial-of-Service-Angriffe beeinflussen die Verfügbarkeit von Systemen und Anwendungen. Um präventiv dagegen vorzugehen, müssen die verdächtigen Anfragen eingedämmt werden. Wenn viele Anfragen in kurzer Zeit von demselben Nutzer ausgehen, ist dies ein Indiz dafür, dass diese keinem normalen Zweck dienen. So kann dieser Nutzer für eine bestimmte Zeit blockiert oder komplett ausgeschlossen werden. Um alle Möglichkeiten von Verwendungen für die Cyber-Sicherheit managen zu können, ist IT-Governance ein wichtiger Aspekt. Dieser wird im folgenden Unterkapitel näher erläutert und analysiert. (Vgl. Beitollahi & Deconinck, 2012: S. 1315 f.)

3.4 IT-Governance im Kontext der Cyber-Sicherheit

In Unternehmensumfeld hat IT-Governance eine Schlüsselposition eingenommen und bringt einen enormen Mehrwert über den IT-Bereich hinaus. IT-Governance verwendet die Informationstechnologie (IT), um Geschäftsziele besser zu erreichen und Risiken angemessen überwachen zu können. Der Einfluss, den die IT auf die Unternehmensperformance ausübt, ist enorm und wird durch die IT-Governance zu einem effektiven Erfolgsfaktor. (Vgl. Raab et al., 2021: S. 1477 f.)

IT-Governance

Die aktuelle IT-Governance steht dem Problem gegenüber, dass die zentralen Mechanismen zu einer unvollständigen und unübersichtlichen Definition führen. Beschrieben werden kann IT-Governance als eine strategiekonforme Steuerung der IT im Sinne des Geschäftszwecks. Die drei Komponenten, Prozesse, Strukturen und Beziehungsmechanismen stellen die tragenden Mechanismen der IT-Governance dar. Die Prozesse können als strategische Entscheidungsfindung und Überwachung in der IT betrachtet werden. Die Strukturen hingegen behandeln den Bestand von zuständigen Funktionen wie IT-Führungskräften und IT-Ausschüssen. Unter dem Begriff Beziehungsmechanismen kann der strategische Dialog und das gemeinsame Lernen von Unternehmen, der Informationstechnologie und Partnerschaften verstanden werden. (Vgl. Raab et al., 2021: S. 1479 f.)

Jede Komponente hat seine Daseinsberechtigung und nur in Kombination von allen drei Bereichen, kann ein ganzheitlicher Ansatz erstellt werden. Dieser fördert eine effektive sowie effiziente IT-Governance in einer Organisation. Es gibt keine Musterlösung für die aufgezeigten Strukturen, Prozesse und Beziehungsmechanismen. Denn jedes Unternehmen muss sich individuell die Mechanismen so kombinieren, dass sie für die Ziele des Unternehmens geeignet sind. Ausschlaggebend für eine geeignete Kombination der Mechanismen können viele Aspekte eines Unternehmens sein. Dazu gehören die Größe, die Branche und auch die Unternehmenskultur, die den Charakter des Unternehmens widerspiegelt. (Vgl. Raab et al., 2021: S. 1480)

Die genannten Mechanismen stehen in Abhängigkeit zueinander, so können gewisse Strukturen ohne die dazu passenden Prozesse, nicht implementier werden. Deshalb muss dabei beachtet werden, dass zentrale Mechanismen koordiniert interagieren. Ein wichtiger Faktor bei den Beziehungsmechanismen ist, dass sie nicht ausschließlich bei der Implementierung der IT-Governance Anklang finden, sondern auch in den Geschäftsbetrieb eingebunden werden. (Vgl. Raab et al., 2021: S. 1480 f.)

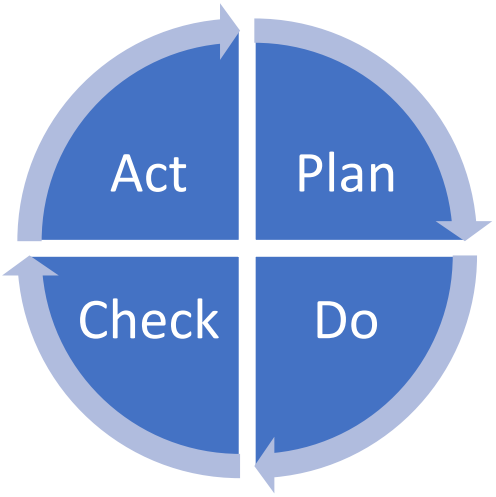

Abbildung 5: P-D-C-A-Zyklus (Quelle: Eigene Darstellung)

IT-Governance kann nicht als ein einmaliger Implementierungsvorgang betrachtet werden. Vielmehr ist es ein eigenständiger Prozess, der in einem Kreislauf dargestellt werden kann. Am besten eignet sich hierfür der P-D-C-A-Zyklus. Dieser Zyklus lässt sich in die vier Phasen, Plan, Do, Check, und Act darstellen, vergleiche Abbildung 5: P-D-C-A-Zyklus. Er kommt zum Einsatz, wenn sich die Problemlösung nicht komplett ergibt. Um an das Ziel zu gelangen, werden Hypothesen formuliert, mit denen der richtige Lösungsweg experimentell gefunden werden soll. Durch das praktische Umsetzen der Hypothese, wird Rückmeldung gegeben, ob das gesetzte Ziel durch diesen Weg erreicht oder die Hypothese angepasst werden muss. (Vgl. Leyendecker & Pötters, 2021: S. 48 f.)

In der Plan-Phase innerhalb des Zyklus werden das Problem und das zu erreichende Ziel in Verbindung gebracht. Es werden Ist- und Zielwerte definiert und miteinander verglichen und analysiert sowie die Problemursachen. Die Methoden, die für die Problemlösung verwendet werden sollen, werden definiert. Zum Ende dieser Phase werden Hypothesen formuliert, die als Lösungsansatz dienen. Die Plan-Phase genießt eine besondere Art der Aufmerksamkeit, da sie die Basis des Zyklus ist, auf dem die folgenden Phasen aufbauen. Deshalb muss die Planung gründlich durchgeführt werden. (Vgl. Leyendecker & Pötters, 2021: S. 49 f.)

Auf die Plan-Phase folgt die Do-Phase. In dieser werden die zuvor erstellten Ansätze umgesetzt, die durch die Hypothese zur Verbesserung beitragen sollen. Abhängig davon welchen Grad an Komplexität das vorliegende Problem besitzt, können die festgelegten Lösungsansätze umgehend verwendet werden. Wenn die Komplexität jedoch zu hoch ist, werden Testläufe durchgeführt. Diese tragen dazu bei, den für diese Komplexität besten Lösungsansatz zu wählen. (Vgl. Leyendecker & Pötters, 2021: S. 51)

Diese werden in der Check-Phase analysiert, indem die erzielten Ergebnisse aus den Tests bewertet werden. Bei Problemen, die während den Testläufen entstehen können, können durch Tests zu den Hypothesen, die entstandenen Ergebnisse überprüft werden. Auf Basis dieser Ergebnisse wird überprüft, ob bezüglich des Problems die Hypothese akzeptiert werden kann oder nicht. (Vgl. Leyendecker & Pötters, 2021: S. 51)

Die letzte Phase des P-D-C-A Zyklus dient dazu, die Resultate aus den anderen Phasen zu analysieren und zu überprüfen, ob und wie weit die festgelegten Ziele erreicht werden konnten. Die Act-Phase beinhaltet das Lernen aus den vorangegangenen Phasen. Wenn die gesetzten Ziele nicht erreicht werden konnten, wird unter Betrachtung der durchgeführten Phasen, die Plan-Phase wieder umgesetzt und der Zyklus beginnt von neuem. Werden die Ziele weitestgehend erreicht, können die verwendeten Lösungsansätze, die Verbesserungen mit sich gebracht haben, innerhalb des Prozesses integriert werden. Darüber hinaus können diese in einem neuen Standard verankert werden, der als Arbeitsgrundlage fungiert, um weitere Probleme zu lösen. Somit wird ein sich wiederholender Zyklus geschaffen, der eine kontinuierliche Verbesserung mit sich bringt. (Vgl. Leyendecker & Pötters, 2021: S. 51)

Cybersecurity Governance

Durch das Kombinieren von IT-Governance und dem P-D-C-A-Zyklus, kann zielgerichtet die Cyber-Sicherheit verstärkt werden. Diesbezüglich kann IT-Governance mit einem ganzheitlichen Konzept, ein Sicherheitsniveau für alle Geschäftsprozesse, Daten als auch IT-Systemen in Organisationen etablieren. Dazu gehören technische als auch nicht technische Sicherheitsanforderungen, die den Einstieg in den Themenbereich Cyber-Sicherheit erleichtern sollen. Damit diese Vorgehensweise funktionieren kann, ist ein Informationssicherheitsmanagement ein wichtiger Bestandteil davon, vergleiche Abbildung 6: Cybersecurity Framework auf Seite 21. Dadurch sollen Vertraulichkeit, Integrität und Verfügbarkeit von Geschäftsprozessen, Anwendungen und IT-Systemen mithilfe eines allgemeinen Risikomanagements gewährleistet werden. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2017: S. 11)

Abbildung 6: Cybersecurity Framework (Quelle: In Anlehnung an National Institute of Standards and Technology)

Für ein konformes Risikomanagement muss eine Beurteilung der Risiken für die Nichteinhaltung der Vertraulichkeit, Integrität und Verfügbarkeit im jeweiligen Anwendungsbereich durchgeführt werden. Dies stellt jedoch einen kontinuierlichen Prozess dar, der ständig überprüft wird und bei Bedarf angepasst und weiterentwickelt wird. Zu beachten sind dabei die Kriterien für das Risiko. Sie müssen zu gültigen und vergleichbaren Ergebnissen führen und das Risiko einem Eigentümer zugeschrieben werden. Dazu gehört außerdem die Abschätzung der Wahrscheinlichkeit, dass die Risiken eintreten sowie die potenziellen Folgen. Den Risiken, denen eine hohe Priorität zugeschrieben werden, sollten zuerst angegangen werden. (Vgl. Deutsches Institut für Normung e. V, 2015: S. 9)

Nachdem die Beurteilung des Risikos stattgefunden hat, muss versucht werden, das Risiko zu beheben oder zu minimieren. Deshalb müssen auf Basis der Risikobeurteilung angemessene Optionen für die Umsetzung ausgewählt werden. Die benötigten Maßnahmen sind nicht vorgeschrieben und können selbst ausgewählt werden. Dazu gehört außerdem, die Wirksamkeit der ausgewählten Maßnahmen zu bewerten. Nur so kann ein Informationssicherheit nach innen und außen gewährleistet werden. (Vgl. Deutsches Institut für Normung e. V, 2015: S. 10)

Ein Sicherheitskonzept wird verwendet, das als Grundlage dient und weiter ausgebaut wird, um neuen Risiken entgegenwirken zu können. Darin werden Sicherheitsmaßnahmen auf ihre Eignung überprüft und daraus gefolgert, ob diese eine Wirkung vorweisen können oder überflüssig sind und abgeschafft oder ausgetauscht werden. Auch gehört der Aspekt dazu, die

Mitarbeiter zu schulen und für diese Themen zu sensibilisieren. Wichtig dabei ist, die Maßnahmen nach ihrer Bedeutsamkeit festzuhalten und zu dokumentieren, damit diese schneller wirksam werden können. Für die verschiedenen Bereiche sollten Verantwortliche ausgewählt werden, die als Ansprechpartner agieren und die Einhaltung von Terminen gewährleisten. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2017: S. 163)

Um einen intakten Informationssicherheitsprozess gewährleisten und verbessern zu können, müssen nicht nur Dokumentationen und die entsprechenden Sicherheitsmaßnahmen in der Geschäftswelt implementiert werden. Vielmehr muss der gesamte Prozess auf allen Ebenen auf die Wirksamkeit und Effizienz überprüft werden. Die daraus entstehenden Ergebnisse und Entscheidungen müssen so dokumentiert werden, dass sie aussagekräftig und verständlich für die spezifischen Zielgruppen sind. Nur so können Fehler sowie Schwachstellen erkannt und behoben werden und dadurch die Effizienz steigern. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2017: S. 164 f.)

Um die Wirksamkeit der Strategie, Maßnahmen und organisatorische Abläufe überprüfen zu können, empfiehlt es sich, Leitaussagen zu definieren, die mit der Hilfe von Kennzahlen analysiert werden können. Diese können bei der Argumentation wertvoll sein und sollten deswegen in Relation mit den festgelegten und erzielten Ergebnissen stehen. Zu beachten dabei ist die Interpretation der Kennzahlen, da diese unterschiedlich betrachtet werden können. Deshalb ist es wichtig, das Ziel und den Aufwand der Messungen vorab zu definieren, um diese als Messfaktor zu verwenden. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2017: S. 165)

Mit einer gut implementierten IT-Governance und der dahinterliegenden Strategie, können Vorgaben für Prozesse, Dokumentationen, Kontrolle als auch Rollen und Verantwortungen auf allen Unternehmensbereichen dazu beitragen, ein einheitliches und dem Risiko entsprechendes Datenschutzniveau zu gewährleisten. Dadurch wird das Risiko für Verletzungen im Datenschutz verringert. Für betroffene Personen einer Datenschutzverletzung wird das Risiko eingedämmt sowie Schäden, Bußgelder und Sanktionen für das Unternehmen verhindert. Außerdem bietet es Transparenz für Außenstehende und eine Grundlage für Audits im Cyber-Sicherheitsbereich. (Vgl. Sowa, 2020: S. 22)

3.5 Zusammenfassung

Cyber-Sicherheit wird immer wichtiger, da Angriffe immer komplexer und raffinierter werden. Dies in Zusammenhang mit der Digitalisierung führt zu Problemen, die Privatpersonen, Unternehmen und Organisationen gleichermaßen betrifft. Die daraus resultierenden Schäden

steigen rasant an aber das Risikobewusstsein verändert sich kaum. Dennoch sind Maßnahmen für die Cyber-Sicherheit unerlässlich, um einen Schutz gegen Angreifer zu besitzen. Dabei müssen sowohl technische Maßnahmen als auch Schulungen durchgeführt werden. So können durch Filter in E-Mails oder durch ein aktives Patchmanagement einige Gefahren abgewendet werden. Dennoch bietet die Schwachstelle Mensch eine große Angriffsfläche, die es zu schützen gilt. Deswegen ist es wichtig, die Mitarbeiter für Themen wie Phishing zu sensibilisieren. Hilfreich für die Kombination von technischen Maßnahmen und der Sensibilisierung der Mitarbeiter ist die IT-Governance mit Bezug auf die Cyber-Sicherheit. Dabei wird der Prozess der IT-Governance mit Cyber-Sicherheit verknüpft, um das Risiko abzuschätzen und auf Ereignisse zu reagieren und aus diesen lernen zu können.

4 Perspektiven der Künstlichen Intelligenz im Kontext der Cyber-Sicherheit

4.1 Kapitelübersicht

Künstliche Intelligenz kann in unterschiedlichsten Bereichen, wie in Unterkapitel 2.5 Anwendungsgebiete beschrieben, implementiert werden. Auch in der Cyber-Sicherheit findet Künstliche Intelligenz anklang und erhält eine immer wichtiger werdende Rolle. Dabei entwickelt sich die Künstliche Intelligenz sowohl als Schutzmaßnahme, als auch als Angriffsmethode stetig weiter und bringt somit Vor- und Nachteile mit sich.

4.2 Adversarial Künstliche Intelligenz

Die Verwendung von Künstlicher Intelligenz setzt immer mehr Anreize, um sie für böswillige Absichten wie beispielsweise Cyber-Angriffe einzusetzen. Auch werden durch Künstliche Intelligenz neue Angriffsvektoren geschaffen, die Schwachstellen beinhalten und so durch Dritte ausgenutzt werden können. Dieses Thema wird immer relevanter, da das bereits stetig steigende Gefährdungspotential durch Cyber-Angriffe dadurch noch weiter ansteigt. Ein Beispiel hierfür sind **Data Poisoning Angriffe**. Diese werden vom Bundesamt für Sicherheit in der Informationstechnik (BSI) wie folgt beschrieben, „durch eine Manipulation der Trainingsdaten des KI-Modells erwirken Angreifer, dass dieses auf (bestimmte) Eingaben nicht wie vom Entwickler vorgesehen reagiert. Aufgrund der vielen Daten und der mangelnden Transparenz sind diese Angriffe meist schwer detektierbar“ (Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 5). Diese Trainingsdaten können nicht nur so manipuliert werden, dass sie Fehler machen. Angreifer können diese mit böswilligen Daten manipulieren, damit im Prozess des Machine Learnings falsche Klassifikationen stattfinden und so Schwachstellen für bestimmte Daten integrieren. Besonders effektiv ist diese Angriffsmethode, wenn die Parameter von dem KI-Modell bekannt sind. Bei dieser Gegebenheit wird die Methode als White-Box-Angriff deklariert. Sind die Parameter für einen Angreifer unbekannt, so wird dieser Angriff als Black-Box-Angriff bezeichnet. Die große Gefahr, die von einem Data Poisoning Angriff ausgeht ist, dass durch das Hinzufügen von gut ausgewählten Mustern des Angreifers, er das Modell für seine ausgewählten Muster kontrolliert. Durch diese zwei Methoden kann einerseits die Verfügbarkeit des Modells bedroht werden. Andererseits kann durch den Angriff die Integrität des Modells nicht mehr gewährleistet werden, da das Modell Entscheidungen trifft, die auf Basis des Angreifers verändert wurden. (Vgl. Bonfanti & Kohler, 2020 S: 2; Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik, Verband der TÜV e.V., 2021: S. 10 ff.)

Die daraus resultierende Gefahr lässt sich bei dem Vorfall von VirusTotal deutlich erkennen. In diesem Fall wurde eine Malware aus einer weit verbreiteten Ransomware-Familie als Start verwendet. Diese wurde dafür genutzt, um mehrere Mutationen von diesem Ransomware-Muster anzufertigen. Diese wurden als Trainingsdaten auf einer Plattform hochgeladen, die für Viren eingerichtet wurde. Auf diese Art wurden die vergifteten Trainingsdaten bereitgestellt und von unterschiedlichen Anbietern verwendet. um diese als Ransomware-Familie zu klassifizieren. Diese konnten jedoch nicht ausgeführt werden. Anhand von diesen Trainingsdaten wurden die Machine Learning Modelle trainiert und für die Identifizierung und Klassifizierung dieser Ransomware-Familie verwendet. Dadurch wird das Gefahrenpotential ersichtlich, da die Ransomware nicht mehr korrekt überprüft werden konnte und so dazu führte, dass diese Ransomware mit einer höheren Wahrscheinlichkeit verwendet werden konnte. (Vgl. Christiaan Beek, 2020)

Evasion Angriffe stellen eine weitere Bedrohung für Künstliche Intelligenz Modelle dar. Dabei sollen Eingabedaten so manipuliert werden, dass das Modell Ausgaben erzeugt, die nicht von den Entwicklern vorgesehen sind. Ein Angreifer verändert durch diesen Vorgang das Modell nicht und somit bleibt es unverändert. Minimale Änderungen an den Eingabedaten können immense Auswirkungen auf die Ausgabe des Modells haben, vergleiche Abbildung 7: Täuschung eines KI-Modells. Selbst Änderungen, die für das menschliche Auge nicht zu erkennen sind oder als irrelevant wahrgenommen werden, können zu Fehleinschätzungen führen. Dabei sind selbst kleine Änderungen von wenigen Pixeln ausschlaggebend und können zu gravierenden Veränderungen der Ausgabe führen. (Vgl. Goodfellow et al., 2014: S. 2 f.)

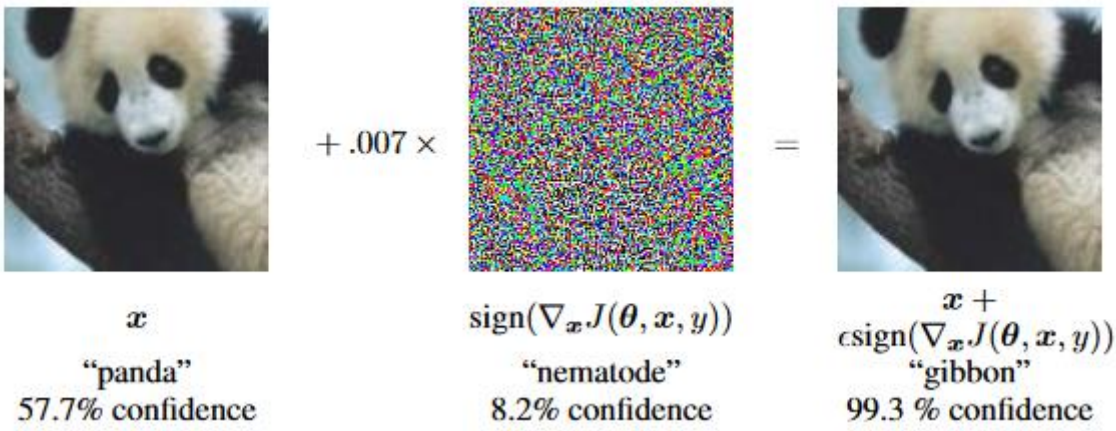

Abbildung 7: Täuschung eines KI-Modells (Quelle: Goodfellow et al., 2014: S. 3)

Besonders deutlich wird die Gefahr im Bereich der Verkehrszeichenerkennung. Dabei können Sticker oder Graffiti auf den Verkehrszeichen platziert werden und führen dadurch zu einer falschen Klassifikation und somit zu einer falschen Ausgabe des Modells. Dies hat zur Folge,

dass Fahrzeuge mit integrierter Verkehrszeichenerkennung Unfälle verursachen können (Vgl. Eykholt et al., 2017: S. 6). Ein Angreifer versucht jedoch, die Veränderungen so gering und unscheinbar wie möglich zu halten, damit diese schwer zu entdecken sind. Dabei ist zu beachten, je komplexer ein Modell ist oder wird, desto einfacher wird es für einen Angreifer, Muster zu finden, die den Ausgabeprozess beeinflussen. Hilfreich für die Stärkung von komplexen KI-Modellen kann die Erhöhung von geeigneten Trainingsdaten sein. Die Stärkung benötigt jedoch mehr Ressourcen, umso größer das KI-System ist. (Vgl. Buchanan, 2020: S. 9; Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 5; Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik, Verband der TÜV e.V., 2021: S. 10 f.)

Die genannten Sticker auf Straßenschilder können des Weiteren im Straßenverkehr gegen autonome Fahrzeuge verwendet werden, vergleiche Abbildung 8: Manipuliertes Verkehrszeichen. Dabei werden die Sticker auf der Straße aufgetragen. Die Erkennung dieser Sticker als Straßenmarkierung führt zu schwerwiegenden Folgen bei autonomen Fahrzeugen. Durch drei platzierte Sticker auf der Fahrbahn, können ein Fahrzeug von der eigentlichen Straßenführung abweichen lassen und auf die Gegenfahrbahn lenken (Vgl. Tencent Keen Security Lab, 2019: S. 33 f.). Dies stellt eine Gefahr für Mensch und Maschine dar. Die Gefahr besteht nicht ausschließlich darin, dass die autonomen Fahrzeuge Fehler machen können. Die Gefahr liegt auch in der Absicherung dieser Systeme, denn ein Angriff auf die Lenkung eins autonomen Fahrzeugs kann möglich sein, wenn Sicherheitsvorkehrungen nicht oder falsch implementiert wurden. So konnte die Lenkung bei einem solchen Fahrzeug während verschiedenen Geschwindigkeiten übernommen und ferngesteuert werden (Vgl. Tencent Keen Security Lab, 2019: S. 13-15).

Abbildung 8: Manipuliertes Verkehrszeichen (Quelle: Eykholt et al., 2017: S. 2)

Des Weiteren gibt es die Angriffsmethode des **Model Stealing Angriffs**. „Angreifer extrahieren die Funktionalität des Modells. Dabei werden Informationen über die Struktur des

Modells, zum Beispiel relevante Entscheidungsparameter, extrahiert oder die Funktionalität des angegriffenen Modells (näherungsweise) kopiert. Ziel ist der Diebstahl geistigen Eigentums oder die Vorbereitung anderer Angriffe“ (Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 5). Da die Entwicklung von ausgereiften KI-Systemen einen enorm hohen Aufwand an Arbeit, Zeit als auch Geld benötigt, kann durch diese Angriffsmethode ein hoher wirtschaftlicher Schaden entstehen. Deswegen werden die Informationen zu den Datensätzen, der Architektur als auch die Parameter des Modells geheim gehalten. (Vgl. Orekondy et al., op. 2019: S. 4949)

Dieser Angriff kann als Blackbox-Angriff durchgeführt werden. Dies bedeutet, dass ohne Kenntnisse über das Entwickelte KI-Modell, Bestandteile von dem Modell extrahiert werden können und in ein eigenes Modell der Künstlichen Intelligenz integriert werden können. Der Angreifer interagiert beispielsweise mit einem ihm unbekannten Modell. Er bietet diesem KI-Modell Bilder zur Eingabe an und erhält dadurch entsprechenden Vorhersagen. Die daraus resultierenden Verbindungen zwischen Bild und Vorhersage, sollen dazu dienen, ein Modell zu trainieren, dass das Verhalten vom Modell des Opfers nachahmt. Somit stehen das Modell des Opfers und das vom Angreifer in Konkurrenz. Dabei können Bilder ohne Kenntnisse der Daten vom originalen Modell verwendet werden. Der Angreifer wählt eine eigene Architektur für sein Modell und versucht durch die Verwendung der eigenen Bilder und den dazu gehörigen Vorhersagen des Opfermodells, so zu trainieren, dass es dieses imitiert. Das Ergebnis einer gelungenen Modell-Extraktion, vergleiche Abbildung 9 Seite 28, konnte eine Genauigkeit des gestohlenen Modells zum Original von 85,7% vorweisen (Vgl. Lekkala et al., 2021: S. 4). Um die Effektivität eines solchen Angriffs zu erhöhen, sollte der Datensatz der Bilder eine Vielzahl an unterschiedlichen Bildern aufweisen. (Vgl. Orekondy et al., op. 2019: S. 4949)

Abbildung 9: Ergebnis einer Modell-Extrahierung für Fußgänger in autonomen Fahrzeugen (Quelle: Lekkala et al., 2021: S. 5)

Bei dem Vorgehen gibt es zwei Varianten, die angewendet werden können. Einerseits kann die randomisierte Strategie angewendet werden. Dabei werden Bilder nach dem Zufallsprinzip genutzt, um Abfragen für das Modell zu machen. Diese Methode führt eine reine Erkundung durch und hat das Problem, dass die verwendeten Bilder, die bereitgestellt werden, irrelevant für den Lernprozess sein können. Die adaptive Strategie hingegen, nutzt ein Belohnungssystem, dass die Abfragen mit Beispielbildern durchführt, die eine höhere Effizienz aufweisen. Abhängig von dem Resultat der Bilder wird eine Belohnung für gute Ergebnisse vergeben, die das Lernen des Modells verstärkt. Dabei wird durch die Belohnung, die verwendete Richtlinie so optimiert, dass die Maximierung der erwarteten Belohnung im Vordergrund steht. Forschungen zeigen, dass durch diese Methode eine Perfomance von 0,82 vom angegriffenen Modell auf das eigene Modell übertragen werden konnte. Somit wird deutlich, welche Gefahr von diesem Angriff ausgeht, denn diese Methode kann für ungefähr $30 durchgeführt werden aber ein sehr viel teureres Modell stehlen. Die erklärte Angriffsmethode wird in Kapitel 6 Fallstudie: Modell-Diebstahl für Zeitreihenprognosen, eine zentrale Rolle einnehmen. (Vgl. Orekondy et al., op. 2019: S. 4951-4956)

Des Weiteren ist zu beachten, dass die Analyse der Arbeitsweise eines Modells ausnutzbare Informationen an Angreifer ergeben können. Die Forscher von Skylight konnten zeigen, dass es ihnen möglich war, einen KI-Malware-Detektor zu täuschen. Sie nutzten öffentliche Informationen über diesen Detektor und konnten dadurch eine Schwachstelle erkennen. Durch diese konnten sie an eine maliziöse Datei eine Zeichenfolge anhängen, die es ermöglichte, das

Bewertungssystem zu täuschen und somit den KI-Malware-Detektor zu umgehen. (Vgl. Adi Ashkenazy & Shahar Zini, 2019)

Angriffe können nicht nur gegen die KI-Modelle gerichtet sein, vergleiche Tabelle 1: Adversarial Künstliche Intelligenz. Die Künstliche Intelligenz selbst kann genutzt werden, um intelligente Angriffe zu starten. Sie kann für Bots verwendet werden, die in sozialen Medien aktiv sind und sich gezielt anfällige Nutzer aussuchen. Sie gehen auf diese ein und beschaffen Informationen. Anhand von diesen werden maliziöse Webseiten, E-Mails oder Links ausgewertet, die eine höhere Wahrscheinlichkeit besitzen, dass ein Opfer auf diese eingeht. Um die Wahrscheinlichkeit der Verwendung noch weiter zu steigern, geben sich die Bots als echte Kontakte des Opfers aus. Sie versenden die Nachrichten über Adressen, die das gleiche Erscheinungsbild wie die der Kontakte besitzen. Dabei imitiert der Bot den Schreibstil des Kontaktes und nutzt menschliches Vertrauen aus, indem er längere Dialoge führt. Somit wird Phishing auf eine neue Ebene gehoben und die Erkennung wird weiter erschwert. (Vgl. Bonfanti & Kohler, 2020: S. 3 f.; Brundage et al., 2018 S. 23 f.)

Die missbräuchliche Verwendung von Identitäten kann durch Künstliche Intelligenz verbessert und dadurch weiterverbreitet werden. Die Verwendung der Identitäten kann für unterschiedlichste Zwecke genutzt werden. Dadurch können Fehlinformationen durch Deepfakes im Internet verbreitet werden. Deepfakes nutzen Künstliche Intelligenz, um Bilder und Videos zusammenzuführen und aus diesen gefälschte aber authentisch wirkende Videos zu generieren. Diese Videos können anschließend so verwendet werden, dass Fehlinformationen durch die darin dargestellte Person im Cyberraum verbreitet werden und dadurch andere Menschen beeinflussen. Des Weiteren können Deepfakes für gezieltes Mobbing, Stalking, Verleumdung oder Phishing verwendet werden. (Vgl. Bonfanti & Kohler, 2020: S. 4; Westerlund, 2019: S. 39)

Tabelle 1: Adversarial Künstliche Intelligenz (Quelle: Eigene Darstellung)

Angriffsart	**Data Poisoning**	**Evasion**	**Model Stealing**	**Angriffe mit KI**
Änderung:	Trainingsdaten des Modells	Eingabedaten des Modells	Keine Änderung	Abhängig vom spezifischen Angriff

Ziele:	Verfügbarkeit und Integrität einschränken, Kontrolle über das Modell für bestimmte Muster	Erzeugung von nicht vorhergesehenen Ausgaben	Diebstahl des Modells	Angriffe effizienter gestalten

4.3 Künstliche Intelligenz für Cyber-Sicherheit

Künstliche Intelligenz kann die Cyber-Sicherheit effektiver gestalten und durch ihren Einfluss, eine unterstützende Wirkung mit sich bringen oder sogar Aufgaben autonom ausführen. Somit stellt die Künstliche Intelligenz einen wichtigen Bestandteil von wirkungsvollen Sicherheitsmodellierungen dar. Dabei nehmen die meisten Modelle sicherheitsrelevante Daten aus unterschiedlichen Quellen, um anhand von diesen Machine Learning zu betreiben. Es werden angefallene Daten aus Anwendungs-, Netzwerk- oder auch Nutzeraktivitäten verwendet, um mit diesen unterschiedliche Ziele verfolgen zu können. Auf der einen Seite kann durch einen Algorithmus die Ausnutzung von Malware festgestellt werden. Dabei stellt Malware einen absichtlich böswilligen in der Software enthaltenen Anhang dar, der auf ein System eines oder mehrerer Opfer aufgespielt wird. Dabei unterscheiden sich die Arten von Malware von ihrem Verhalten und Zielen und können in Virus, Wurm, Trojaner sowie Ransomware unterteilt werden. Andererseits kann die Nutzung so weit reichen, dass riskantes Verhalten festgestellt werden kann. Dieses ist ein Indikator dafür, dass ein Phishing-Angriff möglich ist oder Schadcode eingeschleust werden kann. (Vgl. Aslan & Samet, 2020: S. 6249; Sarker et al., 2021: S. 6 f.)

Für einen aufgabenorientierten Ansatz wird eine besondere Form des Machine Learnings verwendet. Wenn Ziele für Angriffe und Anomalien bekannt sind, können diese in Klassen unterteilt werden. Die Anomalien stellen Abweichungen in Mustern dar, die von einem nicht spezifizierten normalen Verhalten abweichen. Dafür ist die Sonderform des Supervised Learning geeignet. Dabei werden klassifizierte Daten zum Lernen verwendet, um beispielsweise interne Daten von Spam und gefährlichen Aktivitäten innerhalb des Netzwerks unterscheiden zu können. Dafür werden Entscheidungsbäume erstellt oder das Verhalten auf Muster analysiert, aus denen Regeln definiert werden. So können diese Modelle, die vorliegenden Sicherheitsdaten als Basis verwenden und durch ihre Lernfähigkeiten, Malware-

Angriffe klassifizieren und diese vorhersagen. (Vgl. Chandola et al., 2009: S. 2; Pimentel et al., 2014: S. 217; Sarker et al., 2021: S. 7)

Die taktische Ebene der Künstlichen Intelligenz ist hilfreich für die Suche und Analyse von Cyberbedrohungen. Um durch einen Entscheidungsbaum Anomalien und Malware-Angriffe erkennen zu können, wurde das IntruDTree Modell erstellt. Dies wird als Intrusion Detection System (IDS) verwendet. Ein Intrusion Detection System stellt ein Software- oder Hardwaresystem dar, das auftretende Events innerhalb eines Computersystems oder eines Netzwerks aufzeichnet, vergleiche Abbildung 10: Architektur eines KI basierten Intrusion Detection Systems. Diese werden anschließend analysiert und dabei wird darauf geachtet, ob diese die Verfügbarkeit, Integrität oder Vertraulichkeit beeinflussen. Das IntruDTree Modell verwendet Künstliche Intelligenz, um darüber urteilen zu können. Darin wurden 80% der vorliegenden Sicherheitsdaten als Trainingsdaten und die restlichen 20% als Testdaten verwendet. Das daraus entstehende Ergebnis zeigt, dass die vom Modell ungesehenen Testdaten bei der Validierung eine Genauigkeit von 98% aufweisen konnten. (Vgl. Bonfanti & Kohler, 2020: S. 3; Liao et al., 2013: S. 16 f.; Sarker et al., 2020: S. 10 f.)

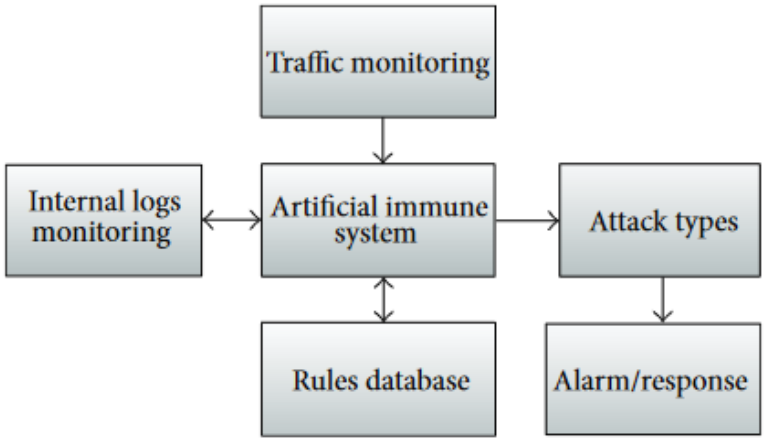

Abbildung 10: Architektur eines KI basierten Intrusion Detection Systems (Quelle: Alrajeh & Lloret, 2013: S. 2)

Die operative Ebene kann genutzt werden, um Daten auszulesen und zu verarbeiten. Durch die Erstellung eines Künstlichen Intelligenz Modells kann mehr Nutzen entstehen als die reine Erkennung von Malware-Angriffen und Anomalien. Heutige Datensätze der Cyber-Sicherheit können eine Vielzahl an Sicherheitsmerkmalen enthalten, die vollständig integriert werden und somit steigt die Komplexität dieser Datensätze. Deshalb ist es wichtig, eine Optimierung bei den Merkmalen durchzuführen, um auf diese Weise die Komplexität eines Sicherheitsmodells zu reduzieren. Richtungsweisend dafür ist, die Überprüfung der Wichtigkeit der einzelnen Sicherheitsmerkmale und deren Auswirkung für die Modellierung. Dadurch finden die

Schlüsselkomponenten mehr Berücksichtigung oder können sich währenddessen herauskristallisieren. Auch können neue Merkmale entdeckt werden, die Einfluss auf das Sicherheitsmodell haben. Durch die erlangten Informationen zu den Merkmalen lassen sich diese anpassen und können bei mangelnder Wichtigkeit entfernt werden. So lassen sich die Komplexität als auch das Sicherheitsmodell an sich vereinfachen. Durch diese Methode konnte mit dem IntruDTree Modell die Genauigkeit von 98% erzielt werden. (Vgl. Bonfanti & Kohler, 2020: S. 3; Sarker et al., 2020: S. 7 ff.; Sarker et al., 2021: S. 8)

Künstliche Intelligenz trägt dazu bei, dass Malware-Angriffe automatisiert erkannt werden können. Jedoch hilft sie darüber hinaus dabei, die Art der Malware zu erkennen. Sie unterscheidet dabei, ob und zu welcher Malware-Familie eine erkannte Malware gehört. Wenn mehrere Malware-Familien für einen Vorfall erkannt werden, werden diese mit der jeweiligen Wahrscheinlichkeit ausgegeben. Falls es zu keiner bekannten Familie zugeordnet werden kann, wird dieser Vorfall als unbekannte Bedrohung deklariert. Dabei kann eine neue Malware-Familie dem Modell angelernt werden, ohne die anderen Funktionen zu beeinflussen. (Vgl. Karbab & Debbabi, 2019: S. 80 f.)

Vorliegende Sicherheitsdaten können in der Realität ungekennzeichnet oder nicht kategorisiert sein. Um diese dennoch verwerten zu können und aus ihnen Muster und Strukturen zu erkennen, wird das Unsupervised Learning verwendet. Dabei besteht die Aufgabe darin, aus nicht gekennzeichneten Daten eine verborgene Struktur innerhalb dieser Daten zu erkennen. Daraus wird eine Funktion generiert, die diese verborgene Struktur beschreibt. Die große Anzahl der enthaltenen Daten verfügen über Merkmale, die als Eingabe integriert sind. Die Ausgabe hingegen wird nicht genauer spezifiziert. Das Modell wertet die integrierten Merkmale aus und kann diese innerhalb von Gruppen oder Cluster kategorisieren. Damit werden ähnliche Merkmale zusammengefasst und so lernt der dahinterliegende Mechanismus des Systems selbstständig, indem er sich an den enthaltenen Informationen und Merkmalen der Daten orientiert und gewöhnt, vergleiche Abbildung 11: Unsupervised Learning auf Seite 33. Im Sicherheitsbereich stellt die Erstellung vom Cluster einen großen Erfolg dar. So können die Sicherheitsdaten mit einer bestimmten Ähnlichkeit berücksichtigt und in Gruppen zusammengefasst werden. Dies kann zur Reaktion auf Vorfälle dienen aber auch für das Risikomanagement, da daraus geregelte Machine Leaning Modelle erstellt werden können. (Vgl. Happiness Ugochi Dike et al., 2018: S. 324f.; Sarker et al., 2021: S. 8)

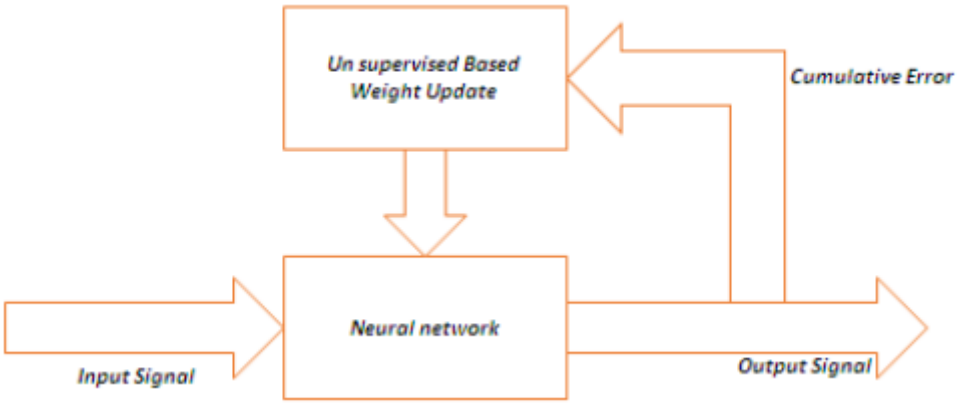

Abbildung 11: Unsupervised Learning (Quelle: Happiness Ugochi Dike et al., 2018: S. 324)

Nicht nur aus datenbasiertem Machine Learning können Anomalien oder Angriffe erkannt werden. Mithilfe von Natural Language Processing (NLP) stehen Möglichkeiten zur Verfügung, um bereits genannte Muster zu erkennen aber es können auch weitere Maßnahmen für andere Sicherheitsvorfälle geschaffen werden. NLP kann als theoriegeleiteten Ansatz verstanden werden, der Computertechniken für automatische Analysen sowie Darstellungen der menschlichen Sprache verwendet (Vgl. Cambria & White, 2014: S. 48). Es wird eine lexikalische Analyse durchgeführt, die sich auf die geregelte Anordnung der Begriffe spezialisiert. Dabei trennt es den gesamten Text nach diesen Anordnungskriterien in Absätze, Sätze oder Schlüsselwörter. Dies kann für die Auswertung und Analyse von Domänennamen verwendet werden, um mit NLP bösartige Domänen zu klassifizieren. (Vgl. Sarker et al., 2021: S. 9)

Um Informationen innerhalb von verschiedenen Netzwerken, Internetzugängen und Hosts zugänglich zu machen, werden Domänennamen verwendet, die einer bestimmten Ressource zugeordnet werden können (Vgl. Mockapetris, 1983: S. 1). Diese werden meist in Phishing-Versuchen angehängt und können mit böswilliger Absicht verwendet werden und durch Ähnlichkeit zu einem vertrauenswürdigen Domänennamen eine Gefahr darstellen. Solche Domänennamen können auch durch Fehler bei der Eingabe entstehen, sodass nicht auf die vertrauenswürdige Domäne zugegriffen wird, vergleiche Abbildung 12: Phishing Beispiel auf Seite 34. Die Gefahr besteht darin, dass sensitiven Daten wie Benutzername, Passwort oder auch Bankinformationen bei dem falschen Domänennamen verwendet werden und so diese Daten an Dritte weitergeleitet werden (Vgl. Buber et al., 2018: S. 609). Um weiter dagegen vorgehen zu können, wird versucht, NLP als Cyber-Sicherheits-Werkzeug zu nutzen, um die Abweichungen zu originalen Domänennamen festzustellen und diese als böswillig identifizieren zu können. Durch diese Erkennung kann die ausgehende Gefahr reduziert werden, indem der Zugang zu diesen Domänennamen eingeschränkt oder blockiert wird. So

konnte mit dem Random Forest Algorithm eine Erfolgsrate von 97,2% erzielt werden und kann so einen Beitrag zu der Erkennung von böswilligen Domänennamen leisten (Vgl. Buber et al., 2018: S. 616 f.).

Der technische Bereich kann bedeutsam für Phishing werden, da in dieser Kategorie, Veränderungen des Verhaltens abweichen können und so nicht dem normalisierten Muster entspricht. Da Phishing aus verschiedenen Kategorien besteht, kann dadurch versucht werden, sich als andere Person auszugeben, um so an sensitive Daten zu gelangen. Dabei kann eine semantische Analyse für die Cyber-Sicherheit an Bedeutung zunehmen. Durch diese Analyse wird auf den Satzbau, die Wahrnehmung der einzelnen Wörter als auch auf das Verständnis des Kontextes eingegangen. Somit kann überprüft werden, ob in Texten, Fragen oder Anweisungen enthalten sind, die zur Preisgabe von sensitiven Daten führen können. Auf diese Weise ist es möglich, Phishing-E-Mails besser entdecken zu können und so die Erfolgsrate eines solchen Angriffs zu reduzieren. (Vgl. Bonfanti & Kohler, 2020: S: 3; Peng et al., 2018: S, 300 f.; Sarker et al., 2021: 9 f.)

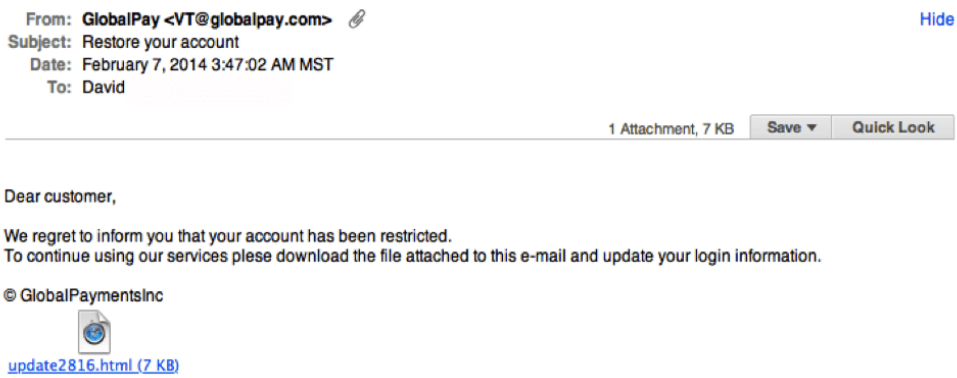

Abbildung 12: Phishing Beispiel (Quelle: SecurityMetrics, 2022)

NLP ist für weitere Bereiche der Cyber-Sicherheit sehr wertvoll. Dazu gehört das Feststellen und Überprüfen von Softwareschwachstellen. Diese können aus Fehlern während der Entwicklung, dem Design oder der Konfiguration der Software hervorgehen. Durch diese Schwachstellen kann die Software ausgenutzt werden und verletzt dadurch Sicherheitsrichtlinien (Vgl. Wu, 2021: S. 5). NLP verwendet als Basis des Lernfortschritts Texte, aus denen die einzelnen Modelle lernen können. Software hingegen wird als Code geschrieben. Deshalb muss NLP den vorhandenen Code als speziellen Text verwerten, um daraus die wichtigen Informationen zu extrahieren, die für die Feststellung von Schwachstellen benötigt werden. Der Vorteil, der sich aus dieser Methode ergibt ist, dass auf diese Weise mehr Muster für das Training und Vorbereitung für weitere Generationen erhalten werden können.

Jedoch liegt zum aktuellen Forschungsstand der Fokus in diesem Bereich auf der Darstellung von Schwachstelleninformationen. Mithilfe von fortgeschrittenen NLP-Modellen können leistungsstarke Funktionen angewandt werden, die wichtige Merkmale für die Bewertung extrahieren können. Wenn diese Informationen über Schwachstellen besser festgestellt werden können, so verbessert sich deutlich die Detektierbarkeit als auch die Beurteilung von Schwachstellen in Software. (Vgl. Bhatti et al., 2021: S. 12 f.; Sarker et al., 2021: S. 10; Wu, 2021: S. 7)

Bei einem Expertensystem übernimmt ein Computerprogramm eine Aufgabe, die eigentlich von einem menschlichen Experten ausgeführt wird. Dabei treffen sie Entscheidungen, die auf Regeln basieren. Diese wurden vorab so definiert, dass sie in dem jeweiligen Themenbereich entsprechen und so Anwendung finden. Das bedeutsamste Merkmal hierbei ist, dass ein Expertensystem schlussfolgern kann, obwohl nicht alle essenziellen Informationen bereitstehen, die für die Schlussfolgerung erforderlich sind. Ein User Interface ermöglicht dabei, die Kommunikation zwischen dem menschlichen Nutzer und dem Expertensystem. Dieses Konstrukt wird in verschiedenen Bereichen eingesetzt. Cyber-Sicherheit stellt einen Verwendungsbereich dar. Durch die Kombination von einem Expertensystem und Künstlicher Intelligenz kann diese Technik verbessert werden. Im Bezug zur Cyber-Sicherheit werden die Regeln und Anforderung für das Expertensystem auf Basis von Sicherheitsrichtlinien hinzugefügt. Anhand dieser wird das Wissen bereitgestellt und die Entscheidungen werden getroffen. Die Künstliche Intelligenz kann dabei unterstützen, die Regeln für das Expertensystem zu optimieren oder neue zu erstellen. Anfragen die häufiger gestellt werden, können besser erkannt werden und die Reaktion darauf wird effizienter. Auch die neuen Regeln tragen dazu bei, ein besseres Cyber-Sicherheits-Expertensystem zu erstellen. (Vgl. Mahdavifar & Ghorbani, 2020: S. 14757 f.; Sarker et al., 2021: S. 12 f.)

4.4 Cyber-Sicherheit für Künstliche Intelligenz

Anwendungen, die Künstliche Intelligenz verwenden, besitzen eigene Anforderungen an die Cyber-Sicherheit. Damit diese Anwendung finden, muss die Cyber-Sicherheit für KI explizit angegangen werden. Da die Anzahl von KI-Anwendungen stetig wächst und immer mehr Aufgaben von ihnen übernommen werden können, wird die Cyber-Sicherheit in Zukunft eine wichtige Rolle einnehmen. Die Fortschritte, die darin erreicht werden können, werden ausschlaggebend sein, wie sicher und widerstandsfähig die KI-Lösungen gegen böswillige Cyber-Aktivtäten, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz, sein werden. Deshalb ist es wichtig, so früh wie möglich Cyber-Sicherheits-Kontrollen für KI-gesteuerte

Anwendungen zu implementieren, die die Integrität der Daten, die Vertraulichkeit sowie die Verfügbarkeit, die Funktion und die Privatsphäre gewährleisten. (Vgl. Executive Office of the President, 2016: S. 36)

Die aktuell erforschten Gegenmaßnahmen für KI-spezifische Angriffe bieten bisher nur einen eingeschränkten Schutz. Dennoch helfen sie dabei, die Angriffsversuche zu erschweren oder die daraus resultierenden Auswirkungen abzumildern. Anhand der Anwendungsfälle muss abgewägt werden, inwieweit diese Verbesserungen ausreichen. Denn jede Anwendung besitzt ein größeres oder kleineres Gefährdungspotential, das die Verwendung von Schutzmaßnahmen erfordert, da Fehlfunktionen und erfolgreiche Angriffe unterschiedliche Auswirkungen haben können. Ein Mindestlevel an Cyber-Sicherheit soll von professionellen Herstellern, Anbieter als auch Entwickler von KI-Systemen gewährleistet werden, indem sie diverse Aspekte berücksichtigen, die im Nachfolgenden beschrieben werden. Herkömmliche Sicherheitsmethoden aus dem Bereich der Software- und Systemsicherheit sollen unverändert in die KI-Systeme integriert werden. Jedoch bietet dies im Themenbereich der Künstlichen Intelligenz keinen ausreichenden Schutz. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 6)

Für ein KI-spezifisches Risikomanagement ist es wichtig, dass der komplette Lebenszyklus für ein KI-System betrachtet wird. Dabei müssen mögliche Risiken erkannt und analysiert werden. Die genannten KI-spezifischen Angriffsmethoden, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz, sollten bei der Implementierung eines Risikomanagements für Künstliche Intelligenz beachtet werden. Mithilfe der Analyse können Maßnahmen abgeleitet werden, die an dem Risiko orientiert, Angriffe und die Auswirkungen abschwächen können. Des Weiteren hilft die Analyse dabei, technische oder organisatorische Maßnahmen zu erkennen, die für die Anpassung der Rahmenbedienungen genutzt werden können. Um beispielsweise ein KI-System gegenüber Adversarial Angriffen zu stärken, kann ein Adversarial Training für das Modell vollzogen werden, um es auf diese Weise zu verstärken. Das Problem hierbei besteht darin, dass durch das Training von ausschließlich einfachen Angriffen, ausgereiftere Angriffe nach wie vor nicht zu verhindern sind. Es empfiehlt sich, adaptive Angriffe auf die eigenen KI-Systeme durchzuführen oder diese von externen geschulten Anbietern ausführen zu lassen. Diese Methode unterstützt dabei, die Risiken valide abschätzen zu können und darüber hinaus helfen sie dabei, die Wirksamkeit der Maßnahmen überprüfen zu können. Um das Risiko neuer Angriffsmethoden auf die eigenen KI-Systeme frühzeitig erkennen zu können, ist es wichtig, die Risikoanalyse regelmäßig durchzuführen und die daraus entstehenden Erkenntnisse umzusetzen. (Vgl. Bundesamt für Sicherheit in der

Informationstechnik, 2021: S. 6; Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik, Verband der TÜV e.V., 2021: S. 13)

Die Qualität von Modellen der Künstlichen Intelligenz ist abhängig von den verwendeten Metriken, die Messwerte darstellen. Deswegen sind diese auch ausschlaggebend für das Gefährdungspotential der Anwendung, da sie die Genauigkeit der erwarteten Eingabedaten beschreibt. Wichtig sind dabei auch Aspekte wie Over-/Underfitting oder Bias-Effekte, die nicht vernachlässigt werden sollten. Bei Overfitting wird das vorliegende KI-Modell mit zu vielen Anpassungen in Bezug auf die Trainingsdaten versehen. Bei Underfitting hingegen sind die Anpassungen nicht ausreichend. Für beide Varianten entstehen dadurch negative Auswirkungen auf die Qualität des jeweiligen Modells. Bias-Effekte können durch die Verwendung von Trainingsdaten mit systematischer Verzerrung entstehen. Dabei sind manche Beziehungen häufiger enthalten als von der Gesellschaft gewünscht oder nicht der Realität entsprechend. Dadurch besteht die Gefahr für das KI-Modell, dass es voreingenommen ist und unfaire Entscheidungen trifft. Deshalb ist es von großem Interesse, dass diese Modelle gegenüber zufälligen oder gezielten Änderungen gestärkt werden. Der beste Ansatz ist es, unterschiedliche Modellansätze anhand von diversen Metriken auf ihre Eignung des Anwendungsgebiets zu überprüfen. Darüber hinaus müssen die Metriken selbst regelmäßig auf die korrekte Funktionsweise und Veränderungen überprüft werden. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 6 f.)

Um KI-Modelle und ihre dazugehörigen Daten gegen Manipulationen zu schützen, ist es erstrebenswert, ein professionelles Datenmanagement zu implementieren. Dabei werden die Trainings- und Testdaten auf ihre Qualität und Quantität überprüft. Des Weiteren ist es essenziell, dass Änderungen an den Daten oder Modellen festgehalten werden und einer Quelle zurechenbar sind. Ein erhöhtes Risiko besteht in der Nutzung von Daten oder Modellen von externen Anbietern, da diese außerhalb des eigenen Einflussbereichs liegen und dadurch willkürlich verändert werden können. Um diese Vorgehensweise im vollen Umfang ausnutzen zu können, ist es wichtig, die erstellten Protokolldaten, die Änderungen beinhalten, auf abweichende Verhaltensmuster zu untersuchen. So können Angriffe auf die Künstliche Intelligenz festgestellt werden und die Nachvollziehbarkeit von diesen Sicherheitsvorfällen ist damit gegeben. Die Erkenntnisse aus den Sicherheitsvorfällen können in einen Prozess integriert werden, der eine schnelle Reaktion auf diese ermöglicht. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 7)

Es können Explainable Artificial Intelligence (XAI) Methoden verwendet werden, um der mangelnden Transparenz und Erklärbarkeit von KI-Modellen entgegenzuwirken. Diese helfen dabei komplexe KI-Modelle durch maschinelle Entscheidungen transparenter und erklärbar zu machen. Dieser Ansatz bietet die Möglichkeit, trainierte Modelle besser verstehen zu können aber in komplexen Systemen kann die Abhängigkeit zwischen Interpretierbarkeit und Vollständigkeit zu Problemen führen. Es können irreführende Erklärungen aus den Problemen entstehen. Dennoch kann es als Unterstützung hinzugezogen werden, um Ergebnisse begrenzt erklären zu können und somit die Fairness von Künstlicher Intelligenz zu verbessern. Grundsätzlich sind aus der Perspektiver der Cyber-Sicherheit einfache und transparente Modelle vorzuziehen. Aus diesem Grund sollten KI-Modelle darauf überprüft werden, wie viel Einfluss welche Parameter besitzt, um so irrelevante Parameter zu entfernen und damit die Komplexität durch Reduzierung der Parameteranzahl zu verringern. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 7; Lossos et al., 2021: S. 312-315)

Damit die KI-Systeme für Verwender transparenter werden, sollten die Anbieter darauf achten, dass sie dem potenziellen Verwender genaue und verständliche Beschreibungen liefern. Diese gehen darauf ein, welche Limitierungen das System besitzt und wie Randbedingungen, die Funktionalität beeinflussen können. So kann der potenzielle Nutzer den Einsatz des Modells für seinen Anwendungsfall selbst entscheiden. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 7)

Die dargestellten Möglichkeiten, um Künstliche Intelligenz absichern zu können, bieten einen beschränkten Schutz für KI-Systeme. Es liegt ein dringender Handlungsbedarf vor, um KI-Systemen einen größeren Schutz bieten zu können. Um Maßstäbe für Sicherheit anbieten zu können, müssen Standards, technische Richtlinien als auch Prüfkriterien und Prüfmethoden entwickelt werden, um eine einheitliche Orientierung zu bieten. Denn zum aktuellen Zeitpunkt gibt es kaum ausreichend geeignete Standards, die diese Möglichkeit bieten. Des Weiteren müssen wirksame Gegenmaßnahmen erforscht werden, die KI-spezifische Angriffe unterbinden. Die existierenden Gegenmaßnahmen sind in den meisten Fällen nicht so weit entwickelt, damit sie den KI-Systemen einen vollkommenen Schutz bieten können. Deshalb müssen die Maßnahmen gegen diese Angriffe weiterentwickelt und weitere mit praktischem Bezug erforscht werden. Defensive Mechanismen, die sich auf typische KI-Angriffe beziehen, werden im weiteren Verlauf dieser Arbeit genauer behandelt, vergleiche Unterkapitel 5.4 Schutzvorkehrungen für KI-Systeme. Ein Beispiel für eine aktuell aufkommende Maßnahme ist AIShield (Vgl. Bosch AIShield, 2022a). Sie versucht weitestgehend KI-Modelle zu schützen. Genauer analysiert wird diese Maßnahme in Unterkapitel 6.4 Gegenmaßnahme

AIShield. Dazugehörend ist auch die Erforschung der Methoden für eine bessere Transparenz und Erklärbarkeit wichtig, da der Mangel daran, die Absicherung der KI-Systemen erschwert. (Vgl. Bundesamt für Sicherheit in der Informationstechnik, 2021: S. 8; Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik, Verband der TÜV e.V., 2021: S. 15 f.)

4.5 Zusammenfassung

Das Potential für Künstliche Intelligenz, diese sowohl für gute Zwecke zu verwenden als auch für böswillige Absichten, ist gegeben. Ein Angreifer, der es auf Künstliche Intelligenz abgesehen hat, kann unterschiedliche Angriffsszenarien durchführen. Dazu gehören Data Poisoning Angriffe, Modell-Diebstahl und Invasion Angriffe. Jedes Angriffsszenario stellt Gefahren für die Künstliche Intelligenz dar, die unterschiedliche Absichten aufweisen. Einerseits kann das KI-Modell gestohlen werden und andererseits die Trainings-, oder Eingabedaten so abgeändert werden, dass dies zu enormen Schäden führen kann. Auch können Angreifer Künstliche Intelligenz für sich nutzen, um Angriffe effizienter zu gestalten und schwerer zu entdecken. Aber auch für die Cyber-Sicherheit lässt sich Künstliche Intelligenz verwenden. Es lassen sich bösartige Inhalte erkennen und auch Klassifizieren oder wenn diese noch nicht bekannt sind, wird eine neue Klassifikation durchgeführt. Darüber hinaus ist es möglich, durch die Erkennung von Anomalien im Inhalt oder der Sprache, interne Daten von bösartigen innerhalb des Netzwerks zu unterscheiden. Des Weiteren lässt sich die Künstliche Intelligenz als Expertensystem einsetzen und kann so selbständig Entscheidungen bezüglich der Sicherheit treffen. Damit die Möglichkeiten, Angriffe auf KI-Modelle auszuüben, reduziert werden kann und dazu beigetragen wird, dass diese vertrauenswürdiger sind, ist Cyber-Sicherheit für Künstliche Intelligenz essenziell. Jedoch ist dieses Thema erst am Anfang und ein vollumfänglicher Schutz ist nicht gegeben. Dennoch lassen sich bereits erste Vorkehrungen treffen. Herkömmliche Sicherheitsvorkehrungen sollten nach wie vor getroffen und ein spezifisches KI-Risikomanagement integriert werden. Dabei muss der gesamte KI-Lebenszyklus abgedeckt sein und wichtige Wertgegenstände, wie die Daten und Metriken geschützt werden.

5 Die Bedeutung von Cyber-Sicherheit Governance in Bezug auf Künstliche Intelligenz

5.1 Kapitelübersicht

Cyber-Sicherheit Governance konnte bereits in den herkömmlichen IT-Anwendungsgebieten dazu beitragen, dass diese sicherer wurden. Somit wurde das Angriffspotential verringert und die Cyber-Sicherheit wurde bei der Entwicklung und Verwendung neuer Technologien berücksichtigt. Deswegen ist es bedeutsam, Künstliche Intelligenz mit Cyber-Sicherheit Governance zu verknüpfen, um damit die Gefahren für dieses Themengebiet erkennen zu können und auf diesen basierend wirkungsvolle Gegenmaßnahmen und Vorkehrungen definieren zu können. Dies ist ein kontinuierlicher Prozess, der sich immer weiterentwickelt, vergleiche Kapitel 3 IT-GRC und Cyber-Sicherheit. Dieser wird für die Zukunft von großer Bedeutung sein, um eine vertrauenswürdige und sichere Künstliche Intelligenz in verschiedensten Anwendungsbereichen voranzutreiben.

5.2 Bedrohungsmodellierung für Künstliche Intelligenz

Um sichere Softwaresysteme, wie auch Künstliche Intelligenz, entwickeln zu können, müssen verschiedenste Herausforderungen bewältigt werden. Durch die Behebung von Sicherheitslücken zu einem frühen Zeitpunkt des Lebenszyklus, beispielsweise im Entwicklungsprozess, führt dazu, dass Kosten für eine Behebung zu einem späteren Zeitpunkt deutlich reduziert werden können. Damit dieser Ansatz verfolgt werden kann, empfiehlt es sich Bedrohungsmodelle systematisch zu analysieren. (Vgl. Hans P. Reiser et al., 2017: S. 7 f.)

Um potenzielle Schwachstellen erkennen und beheben zu können, müssen innerhalb der Bedrohungsmodellierung, die Entwürfe eines Systems methodisch überprüft werden. Diese Vorgehensweise bietet den größten Nutzen, indem innerhalb eines frühen Stadiums des Entwicklungsprozesses eines KI-Systems, die Sicherheitslücken erkannt werden. Die darauf basierenden Erkenntnisse können in entsprechende Schutzmaßnahmen gegenüber den festgestellten Schwachstellen abgeleitet werden. Auf diese Weise kann das ausgehende Risiko eines KI-Modells frühzeitig angegangen und reduziert werden. Risiko untergliedert sich in drei Faktoren, die jeweils Einfluss auf das Risiko nehmen. Zu diesen gehören Bedrohung, Schwachstellen und Auswirkungen. (Vgl. Hans P. Reiser et al., 2017: S. 7 f.)

Bedrohungen lassen sich durch Beeinträchtigungen, Ereignisse oder Umstände beschreiben. Dazu gehören zufällige Fehler und gezielte Angriffe wie Sabotage an den KI-Modellen. Schwachstellen können im Systemdesign, in der Implementierung und im Betrieb des Systems

vorhanden sein. Dennoch ist zu beachten, dass eine Bedrohung oder eine Schwachstelle als einzelner Faktor, keine schädliche Auswirkung zur Folge hat. Damit eine schädliche Auswirkung entstehen kann, muss eine Bedrohung auf eine geeignete Schwachstelle stoßen. Neben den Bedrohungen und Auswirkungen, die ein KI-System ausgesetzt sein kann, ist es wichtig, die Auswirkungen zu betrachten. Dabei können folgende Fragen für die Beurteilung von Auswirkungen hilfreich sein und die Verwendung von Sicherheitsmechanismen erklären. Was macht das KI-System wertvoll? Wie kann dieser bestimmte Wert von anderen Missbraucht werden? Je nach Antwort auf diese Fragen, muss die Entscheidung darüber getroffen werden, ob geeignete Sicherheitsmechanismen vorgenommen werden sollen, um das Risiko zu reduzieren. (Vgl. Hans P. Reiser et al., 2017: S. 8)

Damit Künstliche Intelligenz funktionieren kann, benötigt sie diverse Ressourcen, Umgebungen Prozesse und Benutzer. Diese können je nach Situation und Verwendung zu beabsichtigten oder unabsichtlichen Bedrohungen führen. Dabei unterscheiden sich innerhalb des KI-Zyklus, die Bedrohungslagen jedes Wertgegenstandes, vergleiche Abbildung 13: KI-Wertgegenstände. Dazu gehört jegliche Art der Daten, beispielsweise öffentliche Daten, Trainingsdaten, Testdaten, Rohdaten und viele mehr. Ein weiterer Wertgegenstand für KI-Modelle ist, das Modell selbst. Enthalten sind die verwendeten Algorithmen, Algorithmen des Trainings, Parameter aber auch die Performance des Modells. Eine Bedrohung kann innerhalb der Prozesse entstehen. Verantwortlich kann das Verständnis der Daten, die Datenlagerung oder Modellwartung sein. Des Weiteren gehören Diagramme und Darstellungen der Daten, Zugangslisten sowie die Nutzer als auch die Richtlinien der Künstlichen Intelligenz, zu den schützenden Wertgegenständen. (Vgl. European Union Agency for Cybersecurity, 2020: S. 22 f.)

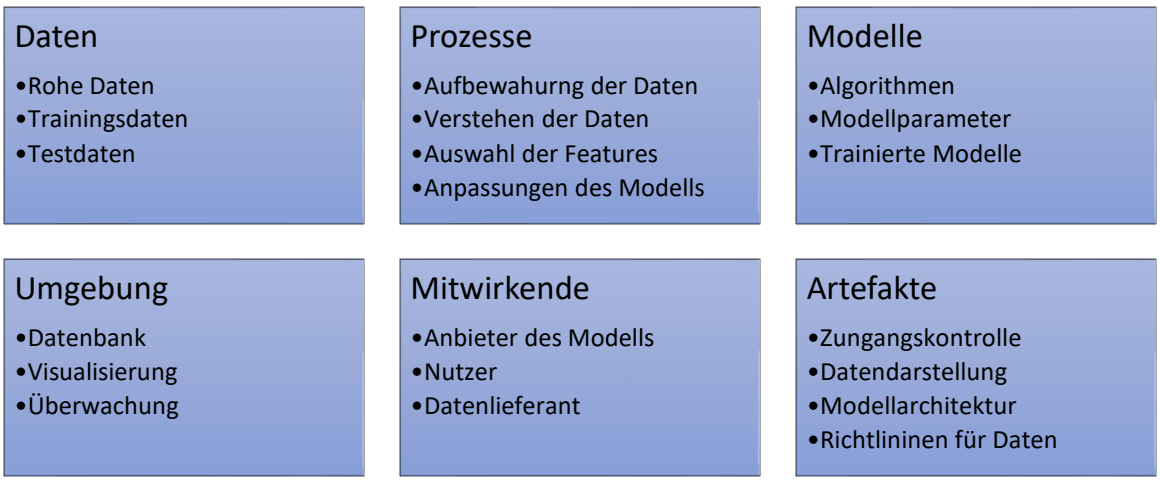

Abbildung 13: KI-Wertgegenstände (Quelle: In Anlehnung an European Union Agency for Cybersecurity, 2020: S. 23)

Künstliche Intelligenz verwendet sehr viele Daten und ist abhängig von diesen. Deshalb müssen diese bei der Bedrohungsmodellierung für Künstliche Intelligenz im Vordergrund stehen. Aus diesem Grund besteht die Bedrohung darin, dass diese Daten entwendet oder verändert werden. Es muss somit darauf geachtet werden, dass der Ursprung der Daten verifiziert und vertrauenswürdig ist. Auch nimmt der Zustand der Daten eine zu schützende Rolle ein, da festgestellt werden muss, welche Änderungen durchgeführt wurden oder ob die Daten sicher gelagert und transferiert werden. Es muss Klarheit über die verschiedenen Methoden, die auf Daten angewandt werden können, herrschen. Dies bezieht sich beispielsweise auf die Normalisierung oder Säuberung der Daten. (Vgl. Ostwald, 2017: S. 2 f.)

Den Daten stehen die Algorithmen und Modelle der Künstlichen Intelligenz gegenüber. Diese besitzen meist einen zu schützenden Status und dazu gehörend sind auch die verwendeten Parameter und die Modellarchitektur. Deshalb ist es für Anbieter und Benutzer von KI-Systemen wichtig zu wissen, ob die daraus resultierenden Ergebnisse korrekt sind. Des Weiteren besteht eine Gefahr für beide Seiten, wenn das Wissen über das Training und die Bewertung fehlen. Wichtig zu wissen in Bezug auf die KI-Modelle und Algorithmen ist, ob aktuelle und repräsentative Daten für diese verwendet wurden und inwieweit diese den Anforderungen und Erwartungen entsprechen. Dies kann zu Gefahren führen, wenn ein Angreifer die Funktionalität eines KI-Modells stiehlt und selbstständig anbietet. Die Vorgehensweise für die Erstellung und den Diebstahl eines KI-Modells werden im nachfolgenden Kapitel genauer beschrieben, vergleiche Kapitel 6 Fallstudie: Modell-Diebstahl für Zeitreihenprognosen. (Vgl. Ostwald, 2017: S. 3)

Dem KI-System selbst, dass die Entscheidungen trifft, stehen Bedrohungen gegenüber. Da für vertrauenswürdige Ergebnisse alle erforderlichen Informationen, dem KI-System bereitstehen müssen, ist darauf zu achten, dass die Verfügbarkeit dieser Informationen nicht durch böswillige oder zufällige Interaktionen unterbunden werden kann. Deshalb müssen Richtlinien zur Absicherung geheim gehalten werden, damit daraus keine Informationen abgeleitet werden können. Auch sollten die erstellten Resultate so dargestellt werden, dass diese unmissverständlich sind und durch potenzielle Erklärungen ergänzt werden. Bestimmte Ergebnisse sollten nachvollziehbar sein, damit die Herkunft bestimmt werden kann. Die Zurückverfolgung der Daten kann eine Bedrohung darstellen, diese Daten können weiterverwendet werden und so Rückschlüsse auf das jeweilige KI-Modell schließen lassen. Deshalb sollte darauf geachtet werden, dass sowohl die eingehenden als auch die ausgehenden Datenströme valide sind, um die Bedrohung von bedrohlichen Abfragen zu erkennen. Auch müssen die Auswirkungen auf ein KI-Modell verstanden werden und wie es die resultierenden

Entscheidungen beeinflusst. Abhängig vom Ziel eines Angreifers, beispielsweise Fehklassifizierung der Daten, Verringerung des Vertrauens oder Manipulation der Trainingsdaten, sind seine Fähigkeiten und seine Kenntnisse über das KI-System für die ausgehende Bedrohung und die dazugehörige Wahrscheinlichkeit ausschlaggebend. (Vgl. Federal Office for Information Security, 2021: S. 25 f.; Ostwald, 2017: S. 3)

5.3 Risk Management Framework

Systeme der Künstlichen Intelligenz können manchmal zu unvorhersehbaren Aktionen fähig sein, da sie ihr Wissen aus Mustern von Daten beziehen und dabei keine Rückschlüsse ziehen, wie diese Muster verursacht werden. Um Künstliche Intelligenz vertrauenswürdig zu machen, müssen bestimmte zentrale Aspekte beachtet werden, um dadurch das Risiko für diesen Themenbereich einschränken zu können. Zu diesen gehören technische Eigenschaften in denen die Genauigkeit, die Verlässlichkeit, die Robustheit und die Resilienz enthalten sind, vergleiche Unterkapitel 4.4 Cyber-Sicherheit für Künstliche Intelligenz. In der sozio-technischen Eigenschaft wird der Fokus auf die Erklärbarkeit, Interpretierbarkeit, Datenschutz sowie Sicherheit und Umgang mit Befangenheit gelegt. In der letzten Kategorie, der Leitsätze, werden die Merkmale Fairness, Verantwortlichkeit und Transparenz in den Vordergrund gestellt. Mithilfe eines Risk Management Frameworks soll dazu beigetragen werden, dass das Vertrauen und die Kommunikation gefördert werden und damit die Risiken der KI-Systeme und deren Umgang besser verstanden werden. Die daraus resultierende bessere Handhabung dieser Systeme soll zur Folge haben, dass Innovationen geschaffen werden können und somit das Potential dieser Technologie voll ausgeschöpft werden kann. (Vgl. Tabassi, Elham, 2022: S. 1)

Viele der Risiken von Künstlicher Intelligenz sind dieselben, wie die anderer IT-Bereiche. Die Risiken, die bisheriger Software oder Systemen gegenüberstehen, gelten auch für das KI-Themengebiet. Deswegen können diese herkömmlichen Risikomanagement Maßnahmen in das Risikomanagement für Künstliche Intelligenz übernommen werden. Jedoch besitzen KI-Systeme einzigartige Herausforderungen, die angegangen werden müssen, um mit diesen Risiken gezielt umgehen zu können. Deswegen soll ein Risk Management Framework dazu dienen, die Lücken der Künstlichen Intelligenz zu schließen, indem es einen strukturierten und dennoch flexiblen Ansatz gegen die Risiken für Unternehmen und die Gesellschaft bietet, vergleiche Abbildung 14: Risikomanagement auf Seite 44. Wichtig dabei ist, dass es sich um einen kontinuierlichen Prozess handelt, der sich mit alten aber auch mit neuen entstehenden Risiken beschäftigt, um diese auf einem akzeptablen Niveau zu halten. (Vgl. Tabassi, Elham, 2022: S. 5)

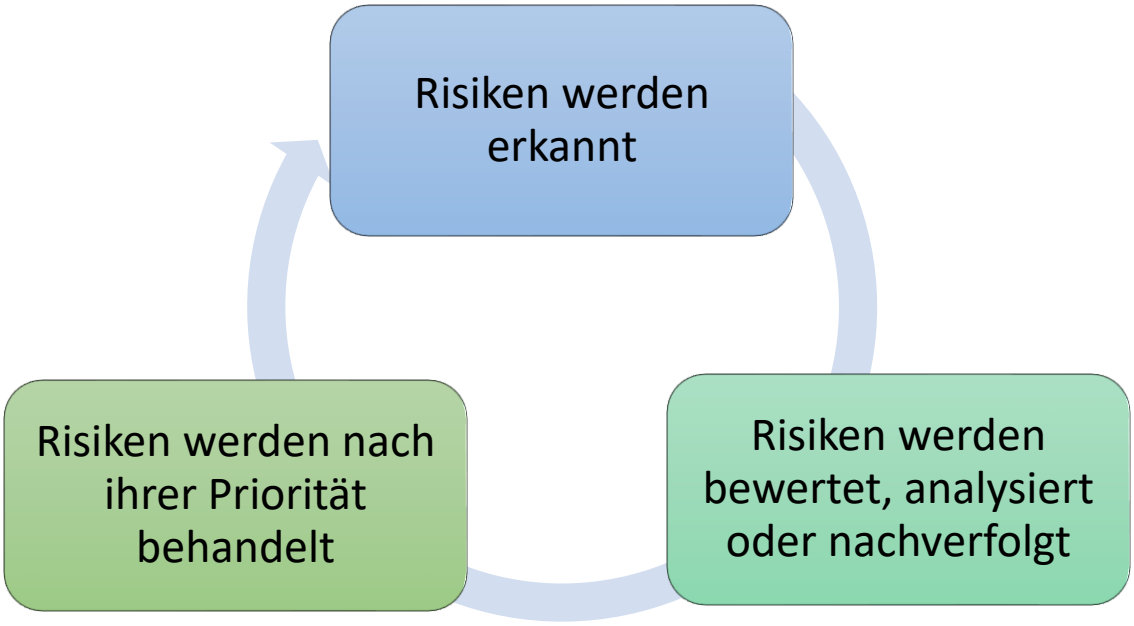

Abbildung 14: Risikomanagement (Quelle: In Anlehnung an Tabassi, Elham, 2022: S. 14)

Risiko beschreibt das Ausmaß, das durch ein mögliches Ereignis entstehen kann. Risiko stellt dabei eine Funktion dar, die negative Auswirkungen eines solchen potenziellen Ereignisses festhält und des Weiteren die dazugehörende Eintrittswahrscheinlichkeit. Betroffen von dem Risiko können Einzelpersonen, Gruppen, Gemeinschaften sowie Systeme, Prozesse und Organisationen sein. Der resultierende Schaden durch KI-Technologien kann je nach betroffener Partei unterschiedlich ausfallen. Wenn Schaden an Menschen angerichtet wird, kann sich dieser in physischen Schaden am individuellen Menschen darstellen oder dessen Rechte und Freiheit werden eingeschränkt. Andererseits können durch Künstliche Intelligenz verschiedene Gruppen an Personen durch das Ergebnis eines KI-Systems diskriminiert werden. Auf gesellschaftlicher Ebene kann die Fairness eines solchen Systems eingeschränkt werden, da dadurch die demokratische Beteiligung unterdrückt werden kann. Für Unternehmen können andere Schäden entstehen. Der Schaden spiegelt sich in Einfluss auf deren technischen Systeme und des Geschäftsbetriebs wider. Dies hat zur Folge, dass die Reputation des Unternehmens leidet, ein finanzieller Schaden entsteht oder die Sicherheit verletzt wird. Schaden an einem System kann sich durch eine weite Vernetzung des Systems zu schwerwiegenden Folgen führen, beispielsweise im Finanzsektor. (Vgl. Wirtz et al., 2022: S. 3 f.)

Quantitative und qualitative Risikomessungen im Bereich Künstlicher Intelligenz sind erschwert, da die Risiken und Auswirkungen noch nicht verständlich genug oder nicht verständlich definiert sind. Ein Problem, das diese Risikomessung weiter erschwert ist, dass KI-Risiken eine zeitliche Dimension besitzen. KI-Risiken variieren während des Lebenszyklus.

Dies hat zur Folge, dass in einer frühen Phase des KI-Lebenszyklus andere Ergebnisse entstehen als zu einem späteren Zeitpunkt. Negative Auswirkungen können hinsichtlich eines kurzen Zeitraums eine geringe Auftrittswahrscheinlichkeit besitzen aber eine hohe zu einem späteren Zeitpunkt. Weitere Risiken sind nicht unmittelbar zu erschließen und können durch die Weiterentwicklung der Künstlichen Intelligenz an Bedeutung gewinnen und müssen im weiteren Prozess erschlossen werden. (Vgl. Tabassi, Elham, 2022: S. 6 f.)

Um das Risiko besser darstellen zu können, werden Schlüssel-Risikoindikatoren verwendet. Diese können aus technischen Faktoren, die für die Bestimmung der Fehlerquoten verwendet werden, bestehen oder auch aus menschlichen Werten. Die einfließenden Faktoren und Werte können anhand von den Auswirkungen und Schäden in Risikostufen wie zum Beispiel niedrig, mittel und hoch eingestuft werden. Wenn jedoch das Risiko durch KI-Systeme zu hoch und unverantwortlich ist, soll nicht nach Maßnahmen gesucht werden, die das Risiko handhaben lassen. Vielmehr muss unter diesen Umständen darauf geachtet werden, ob ein solches KI-System entworfen, entwickelt oder eingesetzt werden sollte. Deshalb ist es wichtig, dass mithilfe von Richtlinien und Normen diese Risikoschwellenwerte durch Regulierungsbehörden, Organisationen oder Branchen festgelegt werden. Die Richtlinien und Normen werden sich vermutlich mit dem Fortschritt der KI weiterentwickeln und werden sich dabei ändern und anpassen. (Vgl. Tabassi, Elham, 2022: S. 6 f.)

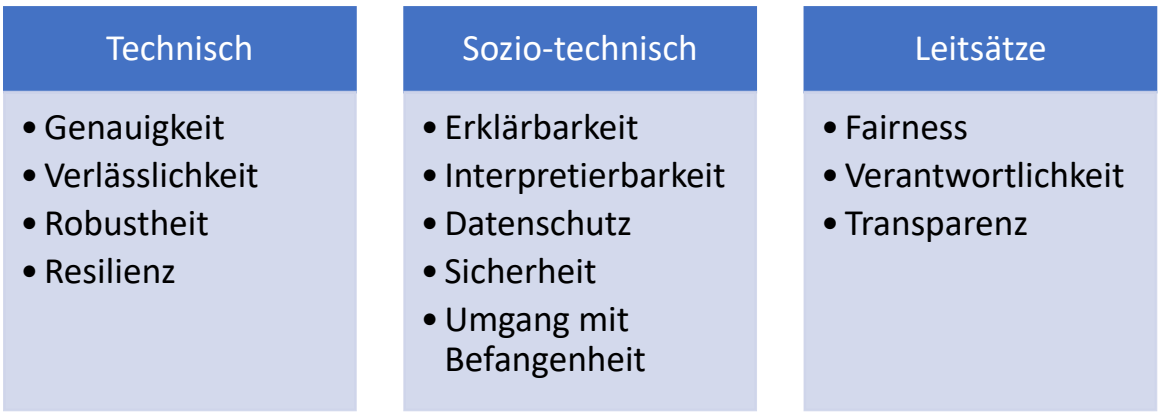

Abbildung 15: KI-Risiken Vertrauenswürdigkeit (Quelle: In Anlehnung an Tabassi, Elham, 2022: S. 8)

Die in Abbildung 15 dargestellten Charakteristiken werden im Nachfolgenden genauer analysiert, um damit ihre Auswirkungen für eine vertrauenswürdige Künstliche Intelligenz zu demonstrieren.

Technische Charakteristiken

Beschreiben lassen sich die technischen Charakteristiken durch die direkte Kontrolle von Systemdesignern und Entwickler auf KI-Systeme. Diese lassen sich durch standardisierte Kriterien evaluieren. Dabei ist die Validität zu beachten. Es sollen keine anderen Dinge von den Daten des Nutzers widergespiegelt werden als ausschließlich die Berechnungsabsichten. Speziell Machine Learning Modelle können durch technische Charakteristiken bewertet und so die Validität überprüft werden. Häufig werden die Aktivitäten von implementierten KI-Systemen durch Überwachung oder Audits auf das beabsichtigte Verhalten überprüft. Es könnte sein, dass bestimmte Berechnungen durch statistische oder Machine Learning Techniken automatisiert werden können. Die Anforderungen, die für Schwellenwerte anfallen, werden als solche spezifiziert. Um die Risiken der Künstlichen Intelligenz anzugehen, können technische Charakteristiken genauer spezifiziert werden. (Vgl. Tabassi, Elham, 2022: S. 8)

Eine essenzielle Rolle nimmt die Genauigkeit ein. Sie beschreibt das Maß der richtigen Abdeckung zwischen einem Machine Learning Modell und den dazugehörigen Trainingsdaten. Dabei haben die Raten von False-Positives und False-Negatives Einfluss auf die Genauigkeit eines KI-Modells sowie die Bewertung von Under-/Overfitting, die bei dem Testprozess zu große Fehlerraten aufweist. Dabei ist in KI Risikomanagement Prozessen zu beachten, dass bei nicht validen KI-Modellen, darauf basierende Entscheidungen getroffen werden, die potenzielle Risiken für Unternehmen und Gesellschaft darstellen. Deswegen ist es grundlegend für die Künstliche Intelligenz, einen Schwellenwert für die Genauigkeit zu definieren, der ausschließlich ein akzeptables Risiko zulässt. (Vgl. Tabassi, Elham, 2022: S. 9)

Verlässlichkeit eines KI-Modells zeichnet sich durch konsistente Resultate aus, die innerhalb eines akzeptablen Fehlerbereichs liegen. Die Verlässlichkeit eines KI-Modells kann durch Techniken, die für die Vermeidung von Overfitting geschaffen wurden, verbessert werden. Darüber hinaus kann auch die Auswahl des Modells entscheiden dafür sein, wie die Verlässlichkeit beeinflusst wird. Die Verlässlichkeit gibt Aufschluss darüber, welches Risiko durch die allgemeine Nutzung und Wiederverwendung der Machine Learning Datensätze oder Modelle ergibt. Zusammen mit der Genauigkeit kann die Verlässlichkeit eine Evaluierung für die Validität für KI-Modelle darstellen. (Vgl. Tabassi, Elham, 2022: S. 9 f.)

Die Empfindlichkeit eines KI-Modells wird als Robustheit dargestellt. Die Robustheit gibt die Reaktion eines KI-Modells an, die durch unkontrollierbare Faktoren entstehen kann. Ein robustes Modell zeichnet sich dadurch aus, dass es trotz vorhandenen Fehlern in den enthaltenen Komponenten weiterhin funktioniert. Währenddessen kann die Performance des

Modells abnehmen oder auf andere Weise angepasst werden, bis die Fehler behoben werden konnten. Robustheit kann die Empfindlichkeit von Ausgaben aber auch von minimal abweichenden Eingaben bemessen. Somit kann die Robustheit, die Empfindlichkeit analysieren und für den Risikomanagement Prozess für Künstliche Intelligenz bereitstellen. (Vgl. Tabassi, Elham, 2022: S. 10)

Resilienz beschreibt den Widerstand, den ein KI-Modell gegen Änderungen der Umgebung oder Nutzung leistet, dazu gehören auch böswillige Angriffe auf das KI-Modell. Resilienz steht in Beziehung mit der Robustheit aber ist weitreichender als die Herkunft der Daten. Sie beschäftigt sich darüber hinaus mit unerwarteter oder böswilliger Nutzung des Modells oder der Daten. Weitere Ansätze im Bezug zur Künstlichen Intelligenz beschäftigen sich mit der Extraktion von Modellen, Trainingsdaten oder anderen intellektuellen Wertgegenständen, die von KI-Systemen gestohlen werden können. (Vgl. Tabassi, Elham, 2022: S. 10)

Sozio-technische Charakteristiken

Sozio-technische Charakteristiken befassen sich mit der Nutzung von KI-Systemen in Bezug auf Individuen, Gruppen und gesellschaftlichen Kontexten. Mentale Repräsentationen von KI-Modellen sollen dabei helfen, die Einhaltung von Richtlinien durch bereitgestellte Ausgaben des jeweiligen KI-Modells zu gewährleisten. Des Weiteren wird darauf geachtet, dass die Operationen, die von einem KI-Modell ausgeführt werden, eine leichte Verständlichkeit aufweisen. Die Ergebnisse, die aus solch einem Modell entstehen, sollen dazu dienen, darauf basierende sinnvolle Entscheidungen zu treffen und zu überprüfen, ob diese auf die Werte der Gesellschaft abgestimmt sind. Die sozio-technischen Faktoren sind im sozialen und organisatorischen Verhalten der Menschheit, innerhalb der Datensätze des Lernprozesses tief verankert. Des Weiteren besteht diese Verbundenheit auch in dem Entwicklungsprozess der KI-Modelle, da in diesem die Entscheidungen der Entwickler miteinfließen. (Vgl. Wirtz et al., 2022: S. 9 f.)

Um sozio-technische Faktoren bemessen und ihre Schwellenwerte definieren zu können, wird zum Zeitpunkt dieser Arbeit, das menschliche Urteilsvermögen eingesetzt. Es ist noch nicht möglich, in Bezug auf sozio-technische Faktoren, das menschliche Urteilsvermögen durch einen automatisierten Prozess zu ersetzten, wie es beispielsweise bei den technischen Charakteristiken möglich ist. Um Risiken der Künstlichen Intelligenz in diesem Anwendungsbereich angehen zu können, ist es wichtig, eine breite und vielfältige Gruppe, bestehend aus verschiedenen Interessensgruppen, für den KI-Lebenszyklus zu bilden. So kann ein Beitrag geleistet werden, um Risiken im Zusammenhang mit den sozialen Aspekten

angemessen handzuhaben, der auf menschlicher Wahrnehmung und Interpretation sowie gesellschaftlichen Werten basiert. (Vgl. Tabassi, Elham, 2022: S. 10)

Die Erklärbarkeit für KI-Modelle nimmt eine wichtige Rolle ein. Sie soll eine Beschreibung darstellen, wie durch ein KI-Modell Entscheidungen erzeugt werden. Obwohl alle Informationen eines KI-Modells verfügbar sind und so eine vollständige Transparenz gewährleistet wird, benötigt ein Mensch technisches Fachwissen. Mit diesem versucht der Mensch, die Funktionsweise des KI-Modells zu verstehen. Die Erklärbarkeit zeichnet sich durch die Wahrnehmung eines Benutzers auf die Funktion eines KI-Modells aus. Beispielsweise wird für eine bestimmte Eingabe, eine zur Eingabe passende Ausgabe erwartet. Erläuterungen befassen sich mit dem Verhalten des KI-Modells oder den Vorhersagen. Diese können hilfreich für die Förderung von Machine Learning oder zur Behebung von Problemen mit dem KI-System oder den Trainingsdaten sein. Auch können diese Erklärungen verwendet werden, um Transparenzanforderungen zu erfüllen. (Vgl. Tabassi, Elham, 2022: S. 11)

Ursachen für eine mangelnde Erklärbarkeit weisen unterschiedliche Faktoren auf. Dazu gehören, eine mangelnde Präzision und Konsistenz der Erklärungen, falsche Ableitungen oder abweichende Funktion des KI-Modells. Die Fähigkeiten eines Nutzers können sich in der Beschreibung der Funktionsweise eines KI-Modells widerspiegeln. Die Erklärbarkeit steht mit der Transparenz in Korrelation. Transparenz ist keine Garantie für Erklärbarkeit, da fehlende Grundlagen eines Nutzers in Machine Learning, die Erklärbarkeit beeinflussen können. Dennoch kann ein transparentes KI-System als erklärbarer betrachtet werden. Ein erklärbares KI-System bringt die Vorteile mit sich, dass es sich leichter testen, überwachen, dokumentieren sowie prüfen und kontrollieren lässt. (Vgl. Tabassi, Elham, 2022: S. 11)

Interpretierbarkeit fokussiert sich auf die Ausgabe eines KI-Modells in Bezug auf den funktionalen Zweck. Erklärbarkeit wird häufig mit Interpretierbarkeit gleichgesetzt aber abweichend zur Interpretierbarkeit, bezieht sich die Erklärbarkeit auf die Funktionsweise des Algorithmus. Die Interpretierbarkeit stellt ein Ausmaß eines KI-Modells dar, das durch die Funktion und der daraus folgenden Resultate, eine Basis für weitere Entscheidungen eines Nutzers bietet. (Vgl. Laurent Dupont, Olivier Fliche, Su Yang, 2020: S. 12)

Um die menschliche Autonomie und die Würde des Menschen zu schützen, ist Privatsphäre essenziell. Sie bezeichnet die Normen und Praktiken, um diese Werte zu schützen und so die Freiheit vor Eindringlingen, Einschränkung der Beobachtung und die Kontrolle der Identität gewährleisten zu können. Deswegen muss darauf geachtet werden, dass keine persönlichen Daten aus dem KI-system entwendet werden können. Jedoch kann die Verwendung von

Künstlicher Intelligenz zu Problemen im Bereich Datenschutz führen, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz. Die Auswirkungen durch datenschutzrelevante Probleme können nicht generalisiert werden, da sie je nach Kultur und Person variieren und so unterschiedliche Ausmaße und Wahrscheinlichkeiten darbieten. (Vgl. Hauschke Andreas & Hildebrandt Stefanie, 2022: S. 24; Tabassi, Elham, 2022: S. 11)

Da KI-Systeme eng mit Menschen zusammenarbeiten, vergleiche Unterkapitel 2.5 Anwendungsgebiete, nimmt Sicherheit eine tragende Rolle ein, da sie mit dem Risiko verbunden ist. Sie ist für die Minimierung von Fehlern verantwortlich, die ein System gefährlich machen können und ist damit ein wichtiger Bestandteil des KI-Risikomanagements. Praktische Ansätze, die durchgeführt werden, um die KI-Sicherheit zu steigern sind, strenge Simulationen, Echtzeitüberwachung oder die schnelle Abschaltung oder Modifizierung von fehlerhaften Systemen. (Vgl. Tabassi, Elham, 2022: S. 12)

Voreingenommenheit ist nicht nur ein Phänomen, das unter Menschen große Auswirkungen haben kann und sich in unterschiedlichsten Formen darstellt. Sowohl positive als auch negative Befangenheit in KI-Systemen können Auswirkungen auf Einzelpersonen, Organisationen und die Gesellschaft ausüben. Diese sind rasanter und haben ein weitaus größeres Ausmaß als die Vorurteile von Menschen. Möglich ist dies, da KI-Technologie enger mit der Gesellschaft verknüpft ist, als die herkömmliche Software. Die Charakteristiken Fairness und Transparenz stehen in engem Kontakt zu der Befangenheit innerhalb der Künstlichen Intelligenz. Diese finden sich in den Leitsätzen wieder (Vgl. Renda, 2019: S. 30 f.; Schwartz et al., 2022: S. 9 f.)

Leitsätze

Es herrscht eine große Einigkeit darüber, dass KI-Technologien anhand von Themengebiet spezifizierten Normen und ethnischen Werten entwickelt werden sollten, obwohl es keinen Standard für ethische Werte gibt. Dies wird möglich, wenn die Normen und Werte in Richtlinien implementiert werden können. So ist es den beteiligten Parteien der Künstlichen Intelligenz möglich, einfache Anforderungen zu formulieren. Die Leitsätze, die für Risiken innerhalb der Künstlichen Intelligenz Betrachtung finden sind, Fairness, Verantwortlichkeit und Transparenz. (Vgl. Tabassi, Elham, 2022: S. 12 f.)

Es gibt kulturelle Unterschiede, die sich in der Wahrnehmung von Fairness widerspiegeln. Dies erschwert es, Standards für Fairness zu definieren. Das Bewusstsein gegenüber Befangenheit von Algorithmen und Datensätzen hat zugenommen und damit wird ein schädliches System durch fehlende Fairness assoziiert. Es gibt viele unterschiedliche Definitionen für Fairness aber

eine Bedingung, die für jede Definition gelten muss ist das Verhindern von schädlichen Befangenheiten, umso Gleichheit und Gerechtigkeit zu stärken und Voreingenommenheit und Diskriminierung zu reduzieren. (Vgl. Tabassi, Elham, 2022: S. 13)

Die Verantwortlichkeit sollte im Falle eines riskanten Ergebnisses vorab geklärt sein. So sollten Organisationen und einzelne Menschen zur Rechenschaft gezogen werden können, für die sie die Verantwortlichkeit übernommen haben. Dazu gehören auch die negativen Auswirkungen, die den Risiken angehören. Auch in Bezug auf die Kultur gibt es abweichende Betrachtungen zwischen der Verantwortlichkeit und dem Risiko. Um verantwortungsvolle KI-Systeme in der Gesellschaft weitestgehend implementieren zu können, ist es wichtig, organisatorische Praktiken zur Schadensbegrenzung wie beispielsweise ein Riskomanagement für Künstliche Intelligenz zu festigen. (Vgl. Tabassi, Elham, 2022: S. 13)

Das Ziel von Transparenz ist es, Informationen zwischen den Betreibern und Verbrauchern von KI-Modellen bereitzustellen. So soll gewährleistet werden, dass der Benutzer ausreichende Informationen für die Interaktion mit einem KI-Modell besitzt. Die bereitzustellenden Informationen sind weitreichend, sie beinhalten Designentscheidungen, Trainingsdaten, Struktur sowie den beabsichtigten Anwendungsfall und Fragen über die Entscheidungen, wie, wann und von wem diese getroffen wurden. Ist die geforderte Transparenz nicht gegeben, so können die Benutzer eines solchen KI-Modells ausschließlich Vermutungen über die fehlenden Informationen aufstellen. Transparenz findet oft Verwendung, um fehlerhafte und benachteiligende Ergebnisse erkennen zu können. Transparenz gewährleistet keinen Datenschutz, Fairness oder ein widerstandsfähiges KI-System. Dennoch kann nicht darüber geurteilt werden, ob ein undurchsichtiges KI-System den genannten Anforderungen ohne Transparenz standhalten kann, während es sich weiterentwickelt und komplexer wird. (Vgl. Federal Office for Information Security, 2021: S. 13; Tabassi, Elham, 2022: S. 13)

5.4 Schutzvorkehrungen für KI-Systeme

Die Erkenntnisse aus der Bedrohungsmodellierung werden verwendet, um angepasste Schutzvorkehrungen für KI-Systeme zu implementieren. Wichtig für die Schutzvorkehrungen bezogen auf die KI-Systeme ist es, die Basis des KI-Modells zu schützen. Dazu gehören die Daten, die für die Entwicklung, Betrieb und die Weiterentwicklung verwendet werden, vergleiche 5.2 Bedrohungsmodellierung für Künstliche Intelligenz. Diese müssen durch ein Identitäts- und Zugriffsmanagement abgesichert werden, um den Zugang für Unbefugte zu verweigern und so die Kompromittierung von Trainings- und Validierungsdatensätzen durch diese zu verhindern. Um die Daten noch weiter zu schützen und die Glaubwürdigkeit dieser zu

steigern, sollte versucht werden, die Daten in verschlüsselter Form abzuspeichern. Darüber hinaus müssen die verwendeten Datenquellen der KI-Systeme dokumentiert werden. Die Dokumentation erfasst alle internen und externen Datenquellen. Dies beschreibt den Zweck der Verwendung und gewährleistet beiläufig die Rückverfolgbarkeit der Daten. Wenn darin Nutzerdaten verarbeitet werden, werden diese in der Systembeschreibung beschrieben. Werden hingegen synthetische Daten für die Erzeugung verwendet, so wird der Prozess dokumentiert und für Nutzer offengelegt. (Vgl. Federal Office for Information Security, 2021: S. 39).

Um Künstliche Intelligenz gegenüber neuen Angriffsvektoren zu schützen und diese auf dem neusten Stand zu halten, sollten KI-Anbieter in quartalsweisem Abstand, ihren Wissenstand hinsichtlich neuer und sich verändernder Angriffsmethoden überprüfen und auffrischen. Ausschließlich das Wissen über neue Bedrohungsszenarien, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz, können dazu beitragen, die KI-Systeme gegenüber diesen Bedrohungen zu schützen. Des Weiteren ist zu beachten, dass Softwarepakete oder Updates Schadcode enthalten können, so sollten diese sorgfältig überprüft werden. (Vgl. Federal Office for Information Security, 2021: S. 26)

Um Schwachstellen innerhalb des KI-Systems erkennen zu können, sollte dieses System durch einen Experten mithilfe von konkreten Angriffen attackiert werden, vergleiche Unterkapitel 5.3 Risk Management Framework. Durch diesen Test lassen sich die eventuell vorhandenen Schwachstellen erkennen und Gegenmaßnahmen angehen. Die Art der Angriffe wird dabei dokumentiert sowie die Testergebnisse und die identifizierten Schwachstellen. Dabei können die Angriffe unterschieden werden in White-Box-Angriffe, Black-Box-Angriffe, übertragbare Angriffe sowie physische Angriffe. Abhängig von der Kritikalität müssen Gegenmaßnahmen implementiert werden. Diese müssen von Fachexperten getestet werden, die nicht am Entwicklungs- und Implementierungsprozess beteiligt waren. Durch diese Tests wird die Wirksamkeit der implementierten Gegenmaßnahmen überprüft. Ein weiterhin bestehendes Restrisiko muss vom Verantwortlichen getragen werden. (Vgl. Federal Office for Information Security, 2021: S. 27)

Um sich gegen Adversarial Angriffe zu schützen, sollten modernste Gegenmaßnahmen implementiert werden, um sich gegen die neusten Angriffe vorzubereiten. Dabei können reaktive Abwehrmaßnahmen wirkungsvoll sein, indem sie auf die Eingabe einwirken, bevor diese das KI-Modell erreichen. Diese Maßnahmen können darin bestehen, die Eingaben auf verdächtiges Verhalten zu überprüfen oder diese vorab zu bearbeiten, damit sie für den Filter des KI-Modells akzeptabel sind. Proaktive Maßnahmen können dazu beitragen, die KI-Modelle

gegen Angriffe abzuhärten. Das KI-Modell kann explizit auf Adversarial Angriffe trainiert werden, damit es diesen standhalten kann. Durch die Verwendung von Generative Adversarial Networks (GAN) kann die Verteidigung gefestigt werden. Diese ermöglichen das Lernen von Repräsentationen, ohne weitgehend kommentierte Trainingsdaten (Vgl. Creswell et al., 2018: S. 53). Um sich gegen Data Poisoning zu schützen, können Sicherheitsmaßnahmen implementiert werden. Dazu gehören die Bereinigung der Daten sowie die Erkennung von Anomalien innerhalb der KI-Datensätze. (Vgl. Federal Office for Information Security, 2021: S. 28)

Anomalien können in KI-Systemen erkannt werden, da bei einer vorgesehenen Nutzung, Datensätze verwendet werden, die keine böswilligen Absichten beinhalten. Dabei sind die Eingabemerkmale definiert und bei gewöhnlichen Daten besteht eine beinahe Normalverteilung. Bei böswilligen Eingaben ist jedoch keine Normalverteilung vorhanden. Anhand dieser Erkenntnis können böswillige und gutwillige Eingaben unterschieden werden, vergleiche Tabelle 2: Verteidigungstaxonomie. Wenn eine böswillige Absicht festgestellt werden kann, ist es möglich, die Anfragen des Angreifers zu unterbinden. Eine andere Option besteht darin, nach der Erkennung die Vorhersagen des KI-Modells abzuändern. Somit kann der Angreifer getäuscht werden und er erhält ein weitaus schlechteres Modell. Jedoch ist es einem Angreifer möglich, eine Normalverteilung innerhalb seiner Eingaben zu erzeugen und so diesen Mechanismus zu umgehen. Deshalb muss dieser Ansatz weiter erforscht werden. (Vgl. Juuti et al., 2018: S. 10-15)

Tabelle 2: Verteidigungstaxonomie (Quelle: Eigene Darstellung)

Angriffsart	**Defensive Maßnahme**	**Details**
Data Poisoning	Erkennung der fremden Daten	Fremde Daten weißen unterschiedliche Merkmale zu den eigenen Daten auf
Modelldiebstahl	Limitierung der Abfragen	Erkennung von auffälligen Anfragen und Reaktion darauf
	Veränderung der Vorhersagen	Verschlechtern der Genauigkeit der Vorhersagen für den Angreifer
Evasion Angriff	Adversarial Training	Fälle für Angriffsfälle erkennen können

5.5 Zusammenfassung

Bedrohungen der Künstlichen Intelligenz können durch Zufall oder durch Angriffe entstehen. Daher muss die Bedrohungslage bekannt sein, um die vorliegenden Bedrohungen so früh wie möglich beheben zu können und so das Risiko zu minimieren. Für eine funktionierende Künstliche Intelligenz ist es bedeutsam, dass alle Informationen zur Verfügung stehen und so eine richtige Entscheidung getroffen werden kann. Besonders wichtig in diesem Themenbereich sind die Daten sowie die verwendeten Algorithmen und ihre Parameter. Damit das Vertrauen in Künstliche Intelligenz durch ein beschränktes Risiko gestärkt werden kann, ist ein Risk Management Framework bedeutend. Innerhalb dieses Prozesses werden die Risiken erkannt, nachverfolgt und nach ihrer Priorität abgearbeitet. Es spielen unterschiedliche Charakteristiken aus dem technischen aber auch sozio-technischen Umfeld eine wichtige Rolle, um das Risiko für Mensch und Maschine auf ein angemessenes Niveau zu reduzieren. Relevant dafür sind Schutzvorkehrungen für KI-Systeme. Es ist sinnvoll, die KI-Systeme auf Schwachstellen überprüfen zu lassen und das Wissen für neuste Bedrohungen regelmäßig aufzufrischen. Des Weiteren sind Maßnahmen wie Zugangsberechtigungen oder verschlüsselte Datensätze wirksam, um diese vor ungewollten Änderungen zu schützen.

6 Fallstudie: Modell-Diebstahl für Zeitreihenprognosen

6.1 Kapitelübersicht

Diese Fallstudie zielt darauf ab, Zeitreihenprognosen und ihre Bedeutung darzustellen, eine Angriffsmethode offenzulegen und darüber hinaus die Verwendung von Boschs AIShield (Vgl. Bosch AIShield, 2022a) als Schutzmaßnahme gegen den genannten Angriff zu diskutieren. Es handelt sich um eine Aufgabe, die noch nicht für den Modell-Diebstahl untersucht wurde, da sich die meisten früheren Arbeiten auf Bilder und Texte beziehen, vergleiche 4.2 Adversarial Künstliche Intelligenz. Dabei wird auf die Entwicklung des KI-Modells und die erlangten Erfahrungen, die während des Prozesses gemacht wurden, eingegangen. Des Weiteren wird das Angriffsszenario anhand des entwickelten Modells dargestellt und analysiert.

6.2 Zeitreihenprognosen

Durch die voranschreitende Digitalisierung haben sich immer mehr IT-Dienste in das private als auch das berufliche Leben integriert. Dabei stehen schnelle und sich ändernde Anforderungen der Digitalisierung gegenüber. Es ergibt sich ein stark schwankender Bedarf an Ressourcen zur Berechnung verschiedenster Anfragen, wie beispielsweise im Cloud Sektor. Um auf diese Lastschwankung reagieren zu können und die benötigte Kapazität rechtzeitig anzupassen, werden proaktive Systeme benötigt, die auf präzisen Prognosemethoden basieren. Wirkungsvoll für diesen Bereich sind Zeitreihenprognosen. Der Ansatz, der hierbei verwendet wird, besteht darin, auf Werte aus der Vergangenheit zurückzugreifen. Diese werden untersucht und basierend auf diesen, Vorhersagen für zukünftige Werte getroffen. Für das Bosch-Geschäft sind diese Zeitreihenprognosen sehr wertvoll, da sie sowohl für Aufgaben im Büro als Umsatzprognosen oder in der Fertigung durch Sensordaten verwendet werden können. Jedoch gibt es keinen universellen Algorithmus, der in jedem möglichen Szenario die bestmögliche Funktionsweise bietet. (Vgl. Hölldobler & Gesellschaft für Informatik e.V., 2021: S. 1 f.)

Die Herausforderung besteht darin, für einen bestehenden Anwendungsfall die optimale Prognosemethode ausfindig zu machen. In den meisten Fällen wird die Prognosemethode nach dem Trial-and-Error Prinzip herausgefunden, da jede Methode Vor- und Nachteile gegenüber dem vorliegenden Anwendungsfall aufweisen kann. Diese Vorgehensweise bringt hohe Kosten mit sich und ist darüber hinaus fehleranfällig. (Vgl. Hölldobler & Gesellschaft für Informatik e.V., 2021: S. 2)

Zeitreihenprognosen haben sich schon lange im Forschungsbereich etabliert und bietet nach wie vor offene Fragen, die bisher noch nicht beantwortet werden konnten. Auch konnte sich

diese Methode in der Produktionsplanung oder auch im Marketing Bereich beweisen und ist in diesen aber auch in vielen weiteren Bereichen unverzichtbar geworden. Die Prognose von zünftigen Daten unterstützen Unternehmen dabei, die Zukunft besser abschätzen zu können und eine bessere Planung und Reaktion in Bezug auf die Zukunft durchzuführen. Auf Basis dieser Informationen können sich Vorteile gegenüber anderen Wettbewerbern ergeben und verbessert oder festigt die aktuelle Position innerhalb des Marktes. (Vgl. Ulrike Fischer, 2013: S. 21)

Früher wurden die Zeitreihenprognosen manuell durch Experten mit langerjähriger Erfahrung durchgeführt. Über die Jahre hinweg und der voranschreitenden Digitalisierung ergaben sich immer mehr Datenquellen, die Daten produzieren. Somit mussten immer größer werdende Datenmengen manuell verarbeitet werden. Dabei ergeben sich nicht nur Herausforderungen, die angegangen werden müssen, sondern auch Möglichkeiten, die noch weitere Erkenntnisse für den jeweiligen Anwendungsbereich liefern können. Dabei lassen sich komplexe Datenbestände in Echtzeit analysieren, indem komplexe statistische Methoden angewandt werden können. Um dieses Prognoseverfahren weiter auszureifen und effizienter zu gestalten, ist es möglich, dieses mithilfe von Künstlicher Intelligenz zu automatisieren. Somit lassen sich bisher ungesehene Beobachtungen anhand von vergangenen Beobachtungen in gewissen Zeitabständen, ohne menschliche Einwirkung erfassen. (Vgl. Ulrike Fischer, 2013: S. 22 f.)

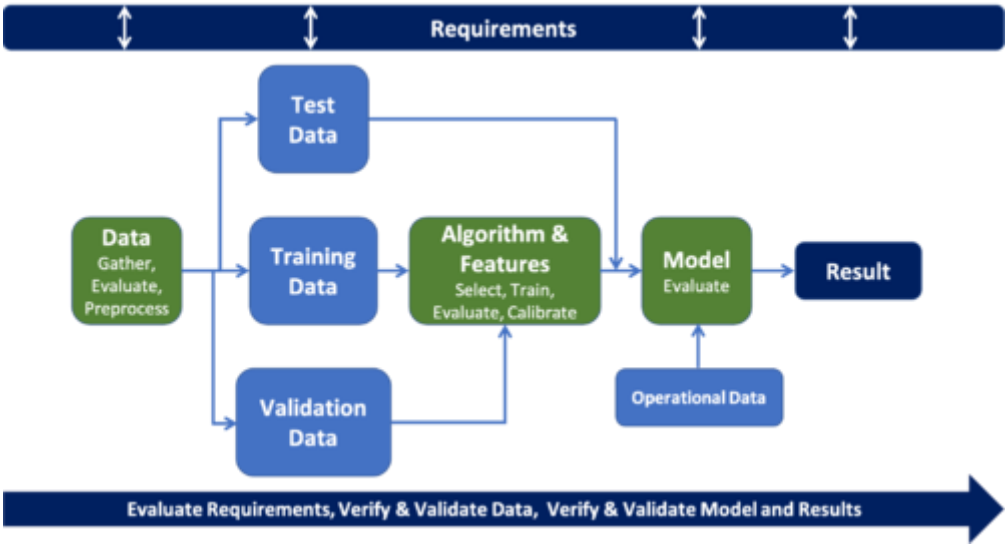

Abbildung 16: Arbeitsweise einer Künstlichen Intelligenz (Quelle: Laura Pullum, 2022)

Damit ein KI-Modell verlässliche Zeitreihenprognosen erstellen kann, muss ein passendes Modell erstellt werden, indem die relevanten Variablen definiert werden und die verschiedenen Parameter abgeschätzt werden können. Damit effiziente und vor allem effektive Parameter verwendet werden können, müssen diese durch ein aufwendiges Optimierungsverfahren ausfindig gemacht werden, vergleiche Abbildung 16: Arbeitsweise einer Künstlichen

Intelligenz. Erst dann kann ein entwickeltes KI-Modell wiederholt und kostengünstig genutzt werden. Deshalb ist es von Bedeutung, das KI-Modell zu schützen. Durch einen implementierten Schutz für Zeitreihenprognosen lassen sich Informationen, die ein Angreifer aus den enthaltenen Daten entwenden kann, verhindern. Ein Angreifer kann darüber hinaus das Ziel haben, die Funktionalität des KI-Modells für sich zu nutzen, indem er es stiehlt. Ein weiteres mögliches Ziel eines Angreifers kann darin bestehen, Daten und Parameter zu verändern, damit falsche Prognosen entstehen und so die Effektivität des KI-Modells zu verändern. Um effektive und verlässliche Zeitreihenprognosen zu erstellen und sich daraus einen Vorteil zu verschaffen, ist es essenziell, das dazugehörende Modell mit ausreichendem Schutz abzusichern. (Vgl. Ulrike Fischer, 2013: S. 22 f.)

KI-Modell für Zeitreihenprognosen

Das Ziel ist, das Angriffsszenario des Modell-Diebstahls anhand eines KI-Modells für Zeitreihenprognosen durchzuführen. Darauffolgend wird die Funktionalität des KI-Modells eines Opfers, durch ein Angreifer KI-Modell versucht, anhand der erlangten Informationen zu imitieren. Zunächst wird ein KI-Modell entwickelt, das Zeitreihenprognosen basierend auf unterschiedlichen Werten, erstellen kann. Danach wird der Modell-Diebstahl an dem KI-Modell durchgeführt und die erlangten Informationen in das Angreifer Modell überführt.

Um ein eigenes KI-Modell zu entwickeln, mussten Datensätze gefunden werden, die für die Verwendung von Zeitreihenprognosen geeignet sind. Die verwendeten Datensätze sind öffentlich freiverfügbar. Für die Entwicklung und Auswertung des KI-Modells wurden zwei Datensätze verwendet, umso die Funktionsweise anhand von unterschiedlichen Daten zu überprüfen. Der erste Datensatz enthält auf dem Wetter basierende Daten (Vgl. Candanedo et al., 2017). Die dazu gehörenden Features sind, die Zeit, die Temperatur und die Luftfeuchtigkeit in verschiedenen Räumen eines Gebäudes sowie die Wetterbedingungen, die außerhalb des Gebäudes herrschten, wie der Luftdruck, Windgeschwindigkeit und Temperatur. Diese Informationen wurden über 4,5 Monate hinweg alle zehn Minuten abgefragt und in den Datensatz integriert, vergleiche Abbildung 17: Ausschnitt der Wetterdaten (Quelle: Eigene Darstellung) auf Seite 57. Die enthaltenen Einträge belaufen sich auf 19735 und besitzen 29 Features. Die Features stellen Merkmale dar, die zu den Prognosen beitragen, abhängig davon, wie viel Einfluss sie in diesem Moment im Vergleich zu den anderen besitzen (Vgl. Wang & Yin, 2021: S. 318 f.).

date	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	...	RH_9	T_out
2016-01-11 17:00:00	60	30	19.89	47.596667	19.2	44.790000	19.79	44.730000	...	45.53	6.600000
2016-01-11 17:10:00	60	30	19.89	46.693333	19.2	44.722500	19.79	44.790000	...	45.56	6.483333
2016-01-11 17:20:00	50	30	19.89	46.300000	19.2	44.626667	19.79	44.933333	...	45.50	6.366667
2016-01-11 17:30:00	50	40	19.89	46.066667	19.2	44.590000	19.79	45.000000	...	45.40	6.250000
2016-01-11 17:40:00	60	40	19.89	46.333333	19.2	44.530000	19.79	45.000000	...	45.40	6.133333

Abbildung 17: Ausschnitt der Wetterdaten (Quelle: Eigene Darstellung)

Dieser Datensatz wurde verwendet, da er die Werte von unterschiedlichen Sensoren enthält und eine Vielzahl an Features besitzt, die für Prognosen verwendet werden können. Des Weiteren wurde er ausgewählt, da er mehr Einträge als der nachfolgende Datensatz beinhaltet, vergleiche Tabelle 3: Vergleich der Datensätze.

date	quarter	department	day	team	targeted_productivity	smv	wip	over_time	incentive	idle_time	idle_men
1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108.0	7080	98	0.0	0
1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NaN	960	0	0.0	0
1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968.0	3660	50	0.0	0
1/1/2015	Quarter1	sweing	Thursday	12	0.80	11.41	968.0	3660	50	0.0	0
1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170.0	1920	50	0.0	0

Abbildung 18: Ausschnitt der Produktivitätsdaten (Quelle: Eigene Darstellung)

Der zweite Datensatz bezieht sich auf die Produktivität eines Unternehmens (Vgl. Imran et al., 2019). Dabei wurden die Informationen über den Tag, die Abteilungen und die Anzahl der Mitarbeiter festgehalten. Des Weiteren wurden die Produktivität, Unterbrechungen sowie unfertige Produkte und Überstunden als Features in den Datensatz integriert, vergleiche Abbildung 18: Ausschnitt der Produktivitätsdaten (Quelle: Eigene Darstellung). Der Datensatz enthält 1197 Werte und 15 Features. Diese Daten wurden im Zeitraum vom 01.01.2015-11.03.2015 täglich für jede Abteilung festgehalten.

Tabelle 3: Vergleich der Datensätze (Quelle: Eigene Darstellung)

Datensatz:	**Wetter**	**Produktivität**
Zeitraum:	11.01.2016-27.05.2016	01.01.2015-11.03.2015
Anzahl der Features:	29	15
Anzahl der Einträge:	19735	1197

Das KI-Modell wurde mit der Programmiersprache Python entwickelt und angegriffen. Dabei wurden unterschiedliche Bibliotheken verwendet, um ein KI-Modell zu implementieren. Zu den verwendeten Bibliotheken gehören TensorFlow, Pandas, Numpy sowie Statsmodels für die Implementierung von ARIMA. Das Vorgehen und die Resultate werden im Folgenden erläutert und dargestellt.

Der erste Ansatz für die Erstellung eines KI-Modells für Zeitreihenprognosen bestand darin, ein ARIMA-Modell zu entwickeln. ARIMA wurde ausgewählt, da es eine effektive und

weitverbreitete Vorgehensweise für Zeitreihenprognosen ist (Vgl. Box et al., 2016: S. 1 f.). Das Wort ARIMA setzt sich aus Autoregressive, Integrated und Moving Average zusammen. Damit sollen stationäre Prozesse vorhergesagt werden. Die Bedeutung von stationär kann in diesem Zusammenhang als keine systematische Veränderung dargestellt werden. Dies ist gegeben, wenn die verwendeten Werte nahe beieinander liegen und nicht zu weit voneinander abweichen. Bei diesem Prozess werden vergangene Werte mithilfe von spezifischen Berechnungsschleifen iterativ angepasst. Dieser Prozess wird solange durchgeführt und angepasst, bis eine Genauigkeit der Vorhersagen getroffen wird, die akzeptabel ist. (Vgl. Jan-Hendrik Meier et al., 2021: S. 161)

Das Modell, das durch ARIMA angelegt werden soll, kann in Teilen oder in vollem Umfang auf Autoregression basieren. Integrated bedeutet für das KI-Modell, dass es nicht auszuschließen ist, dass die Zeitreihe in Serien aufgeteilt werden muss. Jedoch müssen diese Serien in den originalen Zustand zurückversetzt werden, um Prognosen machen zu können. Als letztes ist der Moving Average enthalten, dieser bezeichnet die Durchschnitte der Schätzfehler. Diese stellen die Abweichung zwischen Beobachtungs- und Prognosewerten dar. So kann die Abhängigkeit einer Variablen zu anderen Variablen durch die Regression dargestellt werden. (Vgl. Jan-Hendrik Meier et al., 2021: S. 160 f.)

Die verwendeten Datensätze wurden zunächst mithilfe des Dickey-Fuller-Test auf Stationarität überprüft (Vgl. Dolado et al., 2002: S. 1963 f.). Diese ist ausschlaggeben dafür, damit die Daten für Zeitreihenprognosen verwendet werden können. Nachdem die Stationarität gegeben war, wurde der aktuell verwendete Datensatz in einen Trainings- und Testdatensatz aufgeteilt. Die vorhandenen Daten wurden einem weiteren Test unterzogen, um die Modellierung zu bewerten. Dabei wurde das Akaike-Informationskriterium (AIC) verwendet und daraus die zu verwendenden Parameter abgeleitet, die für die Bestimmung der Muster am besten geeignet sind (Vgl. Richards, 2005: S. 2805). Anhand der Trainingsdaten wurde das ARIMA-Modell mit den bestimmten Parametern trainiert, um daraus die Produktivität des Unternehmens vorhersagen zu können. Die resultierenden Vorhersagen wurden mit den Testdaten evaluiert, vergleiche Abbildung 19: ARIMA-Modell Evaluierung auf Seite 59. Jedoch konnte für dieses KI-Modell nur ein Feature verwendet werden, um darauf basierend Vorhersagen zu treffen. Durch diese Erkenntnis wurde im nächsten Schritt eine andere Modellvariante verwendet.

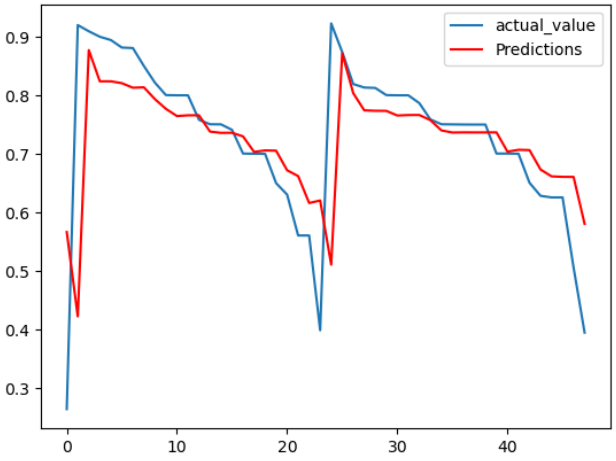

Abbildung 19: ARIMA-Modell Evaluierung (Quelle: Eigene Darstellung)

Datenvorverarbeitung

Der folgende Ansatz konnte das Problem des vorherigen Ansatzes lösen. Im Vergleich zu dem ARIMA-Modell müssen die Daten bei diesem Ansatz vorab bearbeitet werden. Dafür wurden Datenlader implementiert, die den jeweiligen Datensatz für die KI-Anwendung vorbereitet. Diese Datenlader entfernen überflüssige oder nicht verwertbare Informationen aus den Datensätzen. Danach wird der Datensatz in Trainings-, Validierungs-, und Testdaten aufgeteilt. Dieser Schritt ist wie in den beiden Ansätzen zuvor essenziell, da durch die Aufteilung der Daten, das KI-Modell ausschließlich anhand der Trainingsdaten trainiert wird. Ansonsten wird das Ergebnis verfälscht, da die Überprüfung mit bekannten Werten des KI-Modells durchgeführt wird. Der Trainingsanteil stellt den größten Teil mit 70% des gesamten verwendeten Datensatz dar, die Validierung 20% und der Test die restlichen 10%. Im Anschluss werden alle drei Anteile normalisiert und skaliert, um verschiedene Größenordnungen vereinheitlichen zu können, vergleiche Abbildung 20: Normalisierung der Daten auf Seite 60. In diesem KI-Modell werden unterschiedliche Modelle verwendet, die eine bessere oder schlechtere Performance aufweisen. So lässt sich die Effektivität aller Modelle am Ende vergleichen und bewerten.

Um innerhalb von diesem Ansatz, die Daten verwerten zu können, musste ein Windowgenerator entwickelt werden. Für diesen werden Fenster definiert, die anschließend Stichproben aus dem Trainingsdatensatz entnehmen und anhand dieser, das KI-Modell trainiert wird. Die Merkmale, die innerhalb dieses Schrittes Beachtung finden, sind die Breite des Eingabefensters, der zeitliche Abstand und die Verwendung der Features.

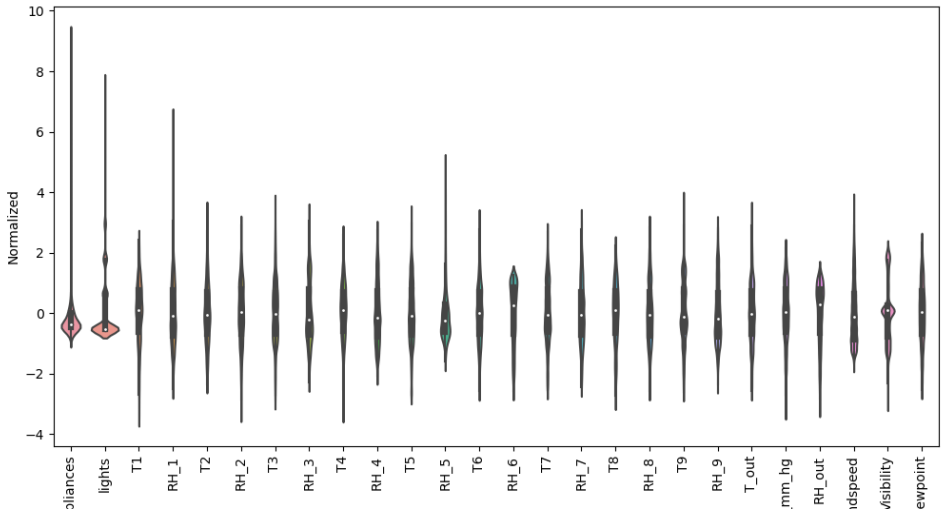

Abbildung 20: Normalisierung der Daten (Quelle: Eigene Darstellung)

Der Windowgenerator verwendet beispielsweise Daten, die vier Stunden in der Vergangenheit liegen, um eine Stunde vorherzusagen. Für dieses Beispiel beträgt die gesamte Größe des Fensters fünf und verwendet vier Werte aus der Vergangenheit, um eine Zukunftsprognose zu erstellen, vergleiche Abbildung 21: Windowgenerator. Dem Windowgenerator werden die Anzahl an Features, die verwendet werden sollen, angegeben und die Anzahl der zu verwendenden Stichproben. Ein Batch legt die Anzahl der zu verwendenden Stichproben fest, bevor das KI-Modell die internen Parameter anpasst. Mit dieser Hilfe werden nach einem Durchlauf eines solchen Batches, die Vorhersagen mit den zu erwartenden Ausgaben verglichen. Die daraus resultierenden Abweichungen werden von dem KI-Modell verwendet, um diese Fehler zu minimieren und auf diese Weise in folgenden Epochen zu lernen. Eine Epoche kann einen oder mehrere Batches enthalten und ist ein Parameter dafür, wie oft der Trainingsdatensatz vom Lernalgorithmus durchgearbeitet werden soll. Damit soll jedem Wert in dem Trainingsdatensatz die Möglichkeit gegeben werden, um die internen Modellparameter zu aktualisieren und so Einfluss auf den Lernprozess zu geben.

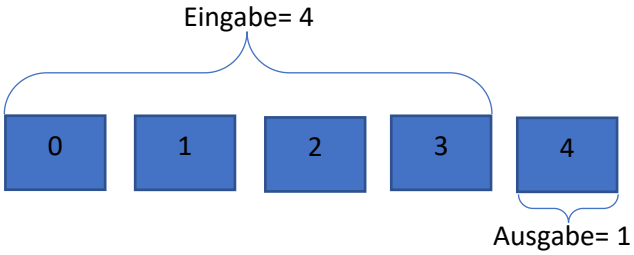

Abbildung 21: Windowgenerator (Quelle: Eigene Darstellung)

Anschließend ist es möglich, den Windowgenerator für verschiedene KI-Modelle zu verwenden. Das erste Modell, das in diesem Ansatz verwendet wurde, ist das Baseline-Modell, es dient zur Überprüfung der Richtigkeit. Dabei verwendet das Modell die aktuelle Eingabe als Vorhersage für die nächste Prognose. Die Ergebnisse dieser Methode kann für komplexere Modelle als Vergleich herangezogen werden, umso die Leistung beurteilen zu können. Diese Vorgehensweise sollte für kurze Prognosen in die Zukunft verwendet werden, da weitfortgeschrittene Vorhersagen geringere Genauigkeit aufweisen können. Die Eingabe wird als Linie dargestellt. Die grünen Kreise auf der Linie stellen die Zielwerte für die Vorhersagen dar und werden zu der Eingabe um einen Schritt verschoben. Sie werden zum Zeitpunkt der Vorhersage und nicht der Eingabe erzeugt. Die Vorhersagen des Modells werden als Kreuze dargestellt und deren Genauigkeit wird darin bemessen, wie weit das jeweilige Kreuz vom Punkt entfernt ist, vergleiche Abbildung 22: Vorhersagen des Baseline-Modells. Die Vorhersagen basieren auf allen enthaltenen Features aber es werden nur Vorhersagen für ein Feature ausgegeben.

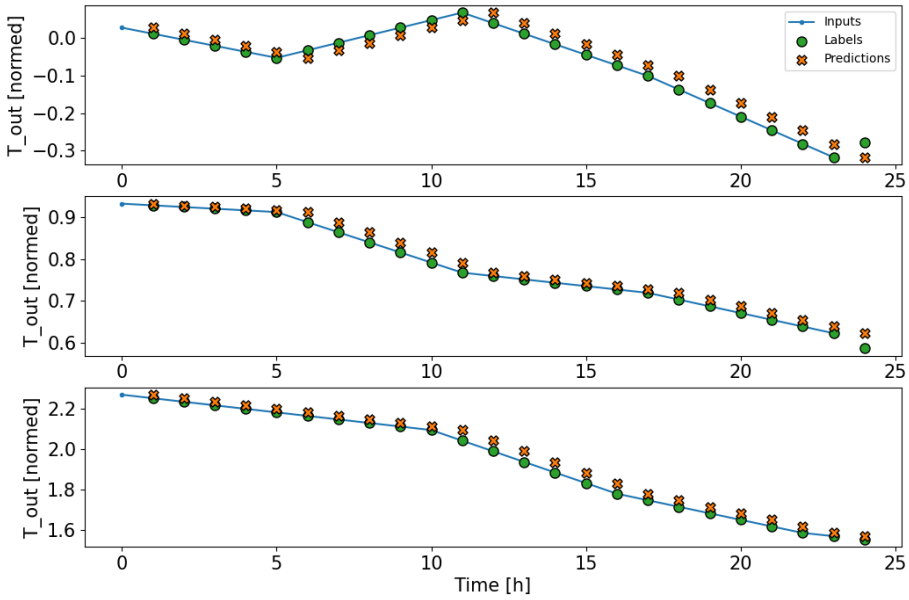

Abbildung 22: Vorhersagen des Baseline-Modells (Quelle: Eigene Darstellung)

Die nachfolgenden Modelle besitzen die Gemeinsamkeit, dass sie neuronale Netzwerke sind. Das nächste Modell, das implementiert wurde, ist das Linear-Modell. Dieses Modell verwendet in jedem Schritt ausschließlich den Eingabewert zu diesem Zeitpunkt. Dem Modell wurde ein Batch von 32 zugeordnet und die Anzahl der Epochen betrug 20. Die Darstellung dieses und der weiteren Modelle geschieht auf die gleiche Weise, wie die des ersten Modells. Der Vorteil des Linear-Modells ist es, dass die Gewichtung der einzelnen Features erkannt und somit

einfach interpretiert werden können, vergleiche Abbildung 23: Vorhersagen und Gewichtung des Linear-Modells.

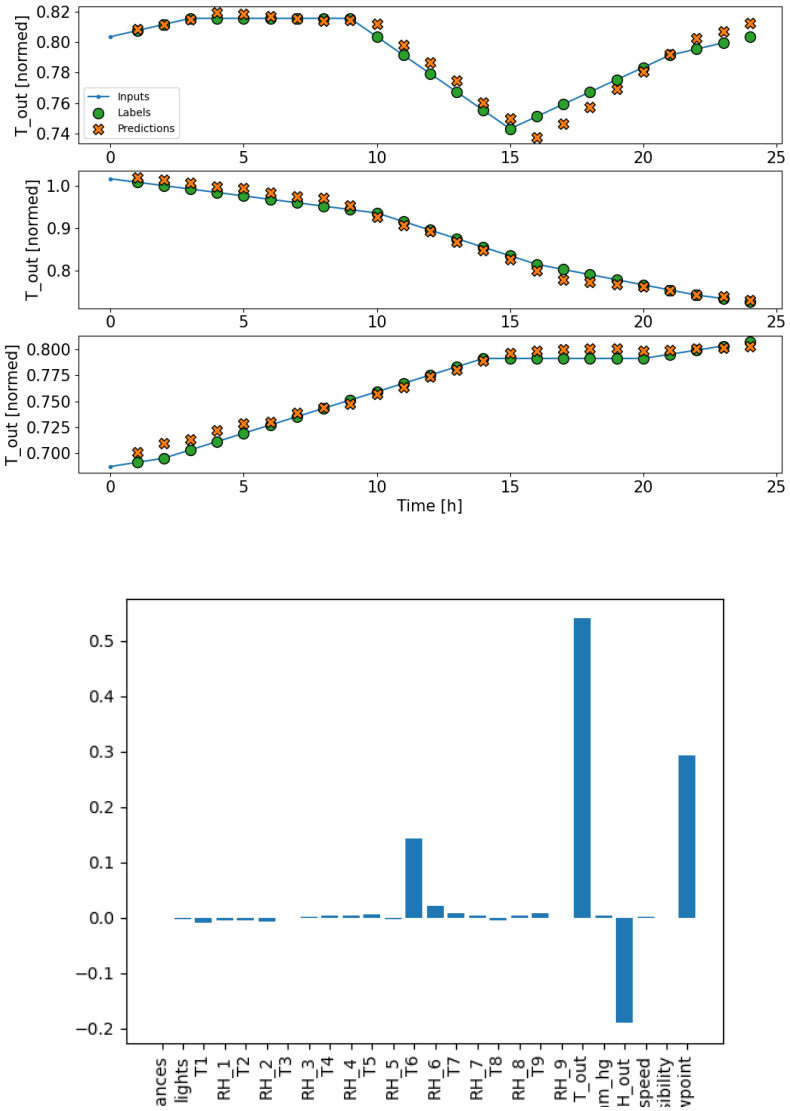

Abbildung 23: Vorhersagen und Gewichtung des Linear-Modells (Quelle: Eigene Darstellung)

Wie die vorherigen beiden Modelle, verwendet das Dense-Modell einzelne Werte. Durch die Erweiterung dieses Modells zum Multi-Step-Dense-Modell, kann es auf verschiedene Zeitschritte zurückgreifen und damit den Zusammenhang dieser Werte erkennen. In diesem Modell wurden drei Zeitschritte verwendet, um einen Zeitschritt vorherzusagen, vergleiche Abbildung 24: Vorhersagen des Multi-Step-Dense-Modells auf Seite 63. Das Problem hierbei ist jedoch, dass das jeweilige Fenster nicht beliebig angepasst werden kann.

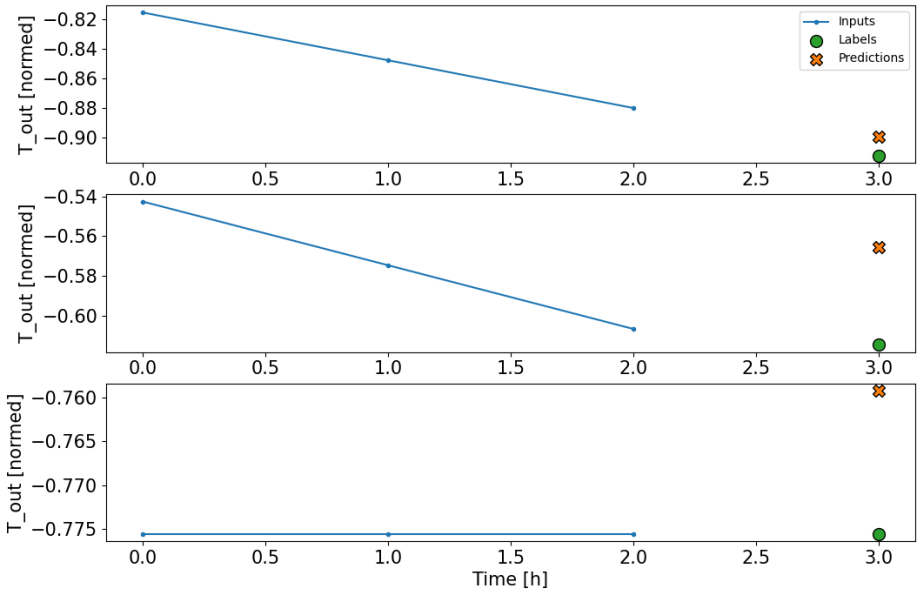

Abbildung 24: Vorhersagen des Multi-Step-Dense-Modells (Quelle: Eigene Darstellung)

Das Problem, das zuvor geschildert wurde, konnte mit dem Convolutional Neural Network (CNN) angegangen werden (Vgl. LeCun et al., 1989). Hierbei werden auch drei Eingaben verwendet, um daraus eine Vorhersage zu treffen. Jedoch unterscheidet sich diese Vorgehensweise zu der vorherigen insoweit, dass eine unterschiedliche Breite des Fensters angegeben werden kann. Das Fenster bewegt sich hierbei über die jeweiligen Eingaben, um daraus die Vorhersagen zu erstellen, vergleiche Abbildung 25: Vorhersagen des CNN. Dabei muss darauf geachtet werden, dass die Anzahl der Werte für die Eingabe größer ist, als die der Ausgabe. Deshalb müssen für die Darstellung des Fensters, weitere Eingaben gegeben werden, damit die Anzahl übereinstimmt.

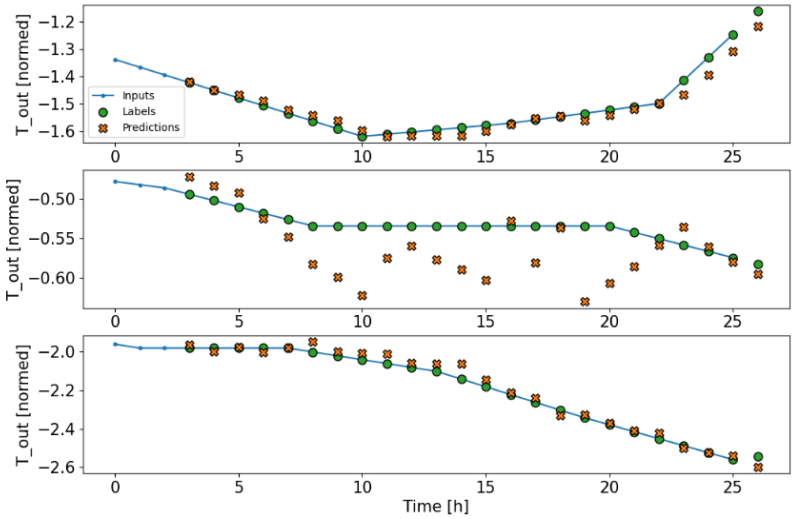

Abbildung 25: Vorhersagen des CNN (Quelle: Eigene Darstellung)

Als Abschluss der verschiedenen Modelle, wurde das Recurrent Neural Network (RNN) (Vgl. Rumelhart et al., 1985) umgesetzt, vergleiche Abbildung 26: Vorhersagen des RNN. Es verwendet Long Short-Term Memory (LSTM), um dieses neuronale Netzwerk zu trainieren (Vgl. Hochreiter & Schmidhuber, 1997: S. 6 ff.). Dabei werden nicht wie bei den anderen Modellen, drei Eingaben zu einer Ausgabe verwertet. Bei dieser Vorgehensweise werden einzelne Eingaben verwendet aber diese werden innerhalb des Modells einzeln abgearbeitet und jeweils in einem internen Zustand festgehalten, umso Vorhersagen zu treffen (Vgl. van Houdt et al., 2020: S. 5931 f.).

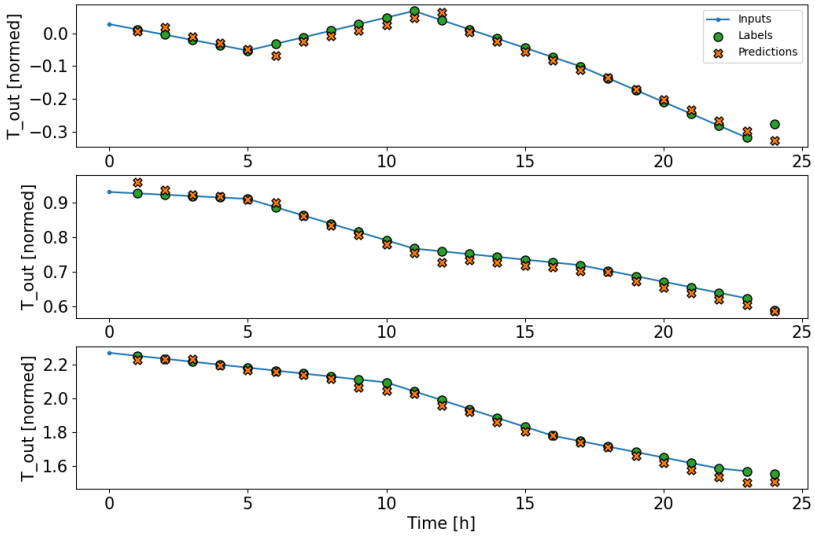

Abbildung 26: Vorhersagen des RNN (Quelle: Eigene Darstellung)

Jedes Modell bringt eigene Vorteile mit sich und kann je nach Datensatz eine bessere oder schlechtere Performance aufweisen. Für alle Modelle, die Epochen durchlaufen, wurden diese auf 20 und die Größe der Batches auf 32 festgelegt. Je geringer der mittlere absolute Fehler des einzelnen Modells ist, desto besser ist die Performance. Dabei stellt der mittlere absolute Fehler ein Maß für die Abweichung dar (Vgl. Karunasingha, 2022: S. 610 f.). Diese Methode wird für alle Auswertungen in dieser Arbeit verwendet. Unter den Modellen, die ausschließlich eine Eingabe für die Ausgabe verwenden, schneidet das Baseline-Modell am besten ab. Die beste Performance unter den Modellen, die mehrere Eingaben für eine Ausgabe verwenden, liefert das CNN. Dies wurde durch Conv abgekürzt und besitzt den niedrigsten mittleren absoluten Fehler in dieser Teilmenge, vergleiche Abbildung 27: Performance der Modelle auf Seite 65.

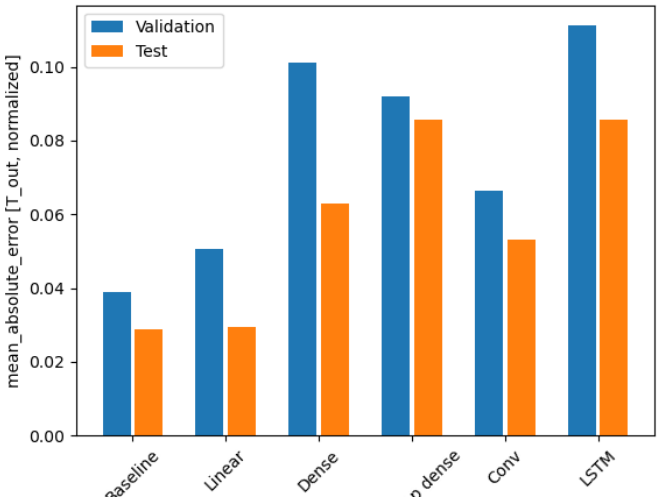

Abbildung 27: Performance der Modelle (Quelle: Eigene Darstellung)

Nachfolgend werden die verwendeten KI-Modelle angegriffen und versucht, das KI-Modell anhand von Prognosen zu stehlen. Dabei werden die unterschiedlichen Prognosen der einzelnen KI-Modelle verwendet, um dasselbe und die anderen KI-Modelle zu trainieren und so ein Angreifer Modell zu erzeugen. Dabei werden die Auswirkungen der unterschiedlichen Prognosen, beispielsweise die des CNN, für das Angreifer Modell eine wichtige Bedeutung haben.

6.3 Modell-Diebstahl

In diesem Unterkapitel wird versucht, die Funktionalität des zuvor erstellten KI-Modells zu entwenden, um somit zu zeigen, wie schnell ein solcher Modelldiebstahl durchgeführt werden kann und dadurch große wirtschaftliche Auswirkungen für Unternehmen entstehen können, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz. Damit der Angriff ein realistischeres Szenario erhält, wird das Feature, das vom Angreifer später vorhergesagt werden soll, entfernt. Für diesen Angriff wurden randomisierte Werte anhand der Gaußschen Normalverteilung erstellt und innerhalb eines Datensatzes abgespeichert. Dieses Vorgehen verwendet ein Angreifer, wenn dieser keinen Zugang zu dem Datensatz besitzt, an dem das Modell trainiert wurde. Dabei wurden die erzeugten Werte in 25 Features festgehalten. Dieser Datensatz wird verwendet, um aus diesen Daten, Prognosen zu erhalten.

Der Datensatz des Angreifers wird anstelle der originalen Daten des Modells als Testdaten verwendet. So nutzt das KI-Modell des Opfers, die Angreifer Daten und erstellt anhand von diesen, Prognosen für die verschiedenen verwendeten Modelle, die darin implementiert wurden. Die Prognosen für das entfernte Feature werden gespeichert und mit den Angreifer Daten zusammengeführt, sodass dieses Feature den Datensatz erweitert. Darauffolgend wird

dieser Datensatz verwendet, um ein Angreifer KI-Modell zu trainieren. In diesem Fall besitzt das Angreifer KI-Modell, dieselbe Struktur und Aufbau, wie das KI-Modell des Opfers.

Indem Angreifer Modell werden die entwendeten Vorhersagen von drei integrierten Modellen verwendet, da sowohl das Dense Modell als auch das Multi Step Dense Modell ausschließlich eine Prognose in die Zukunft erzeugen können, wurden der Fokus auf die anderen drei Modelle gelegt. Das Baseline Modell wird hierbei nicht beachtet, da es kein trainiertes Modell darstellt. Genauer analysiert werden die Prognosen des Linear-, Convolutional-, und des Recurrent Modells. Je nach Modell weisen die Resultate eine unterschiedliche Genauigkeit auf. Das beste Ergebnis im Vergleich zum Original konnte das CNN mit den dazugehörenden Prognosen erzielen, vergleiche Tabelle 4: Mittlerer absoluter Fehler des Angreifer Modells.

Tabelle 4: Mittlerer absoluter Fehler des Angreifer Modells (Quelle: Eigene Darstellung)

Angreifer/Opfer	*Original*	*Linear*	*Convolutional*	*Recurrent*
Linear	0,0982	1,0361	0,2066	0,5457
Convolutional	0,0986	0,1779	0,1045	0,1508
Recurrent	0,0935	0,3058	0,1116	0,2794

Die Abweichungen der mittleren absoluten Fehler, die für die unterschiedlichen Modelle dabei erzielt werden konnten, kommen der Genauigkeit vom originalen Modell bedeutend nahe. Somit ist es möglich, je nach verwendetem Modell, das Modell des Opfers besser oder schlechter zu imitieren und kann so ein Großteil des Entwicklungsprozesses einsparen. Das Training des Angreifer Modells konnte in wenigen Minuten durchgeführt werden. Dies bedeutet für ein Unternehmen, dass die gesamte Entwicklung eines solchen KI-Systems, von einem Angreifer innerhalb von kurzer Zeit umgangen werden kann und so ein funktionsfähiges und fast gleich performendes KI-Modell für Zeitreihenprognosen erhält, vergleiche Abbildung 28: Angreifer Prognosen des CNN auf Seite 67. Dabei können selbst Prognosen aus anderen KI-Modellen für sich unterscheidende KI-Modelle verwendet werden. Dadurch entstehen für einen Angreifer kaum vorhandene Kosten, die jedoch für das Opfer sehr hoch sein können, durch den Entwicklungsprozess. Um dies zu verhindern und die KI-Systeme zu schützen, ist es unabdingbar, dafür Gegenmaßnahmen zu implementieren.

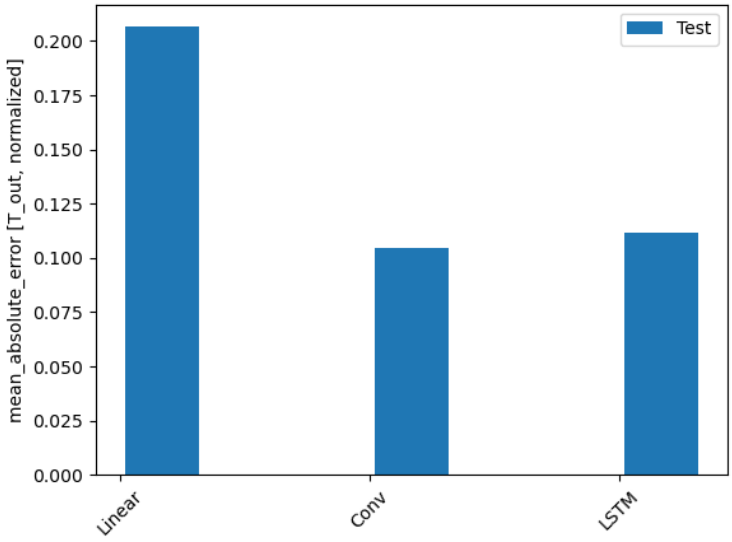

Abbildung 28: Angreifer Prognosen des CNN (Quelle: Eigene Darstellung)

6.4 Gegenmaßnahme AIShield

Durch das langsam steigende Bewusstsein gegenüber Angriffen auf Künstliche Intelligenz, und den dahinterliegenden Modellen und Daten, entstehen immer wieder neue Ansätze, diese vor solchen Angriffen zu schützen. Eine Gegenmaßnahme, die im Januar 2022 angekündigt wurde, ist AIShield. Eine große Herausforderung für die Absicherung von KI-Systemen ist, das Forschungswissen über neue Angriffe und die zu verwendenden Maßnahmen gegen diese. AIShield unterstützt dabei, indem es Schwachstellen der KI-Anwendung bewertet und eine Sicherheitshärtung durchführt. Dabei greift AIShield auf eine Angriffsdatenbank zurück, die stetig aktualisiert wird. Die Bewertung wird anhand von relevanten Angriffsvektoren durchgeführt. Die daraus resultierenden Werte dienen als Basis für die Schaffung eines Abwehrmechanismus gegen Angriffe. Dieser passt sich den unterschiedlichen KI-Modellen oder verwendeten Datensätzen an. (Vgl. Manoj Parmar & Amit Phadke, 2022)

Darüber hinaus bietet es Benachrichtigungen über Angriffe, sobald diese durchgeführt werden. Diese können in verschiedene Management-Services implementiert werden. Die bereitgestellte Benutzeroberfläche stellt informative Berichte und Visualisierungen zur Verfügung, die weiteren Aufschluss über die Angriffe liefern. Des Weiteren ist AIShield skalierbar. So kann es für kleine und große Projekte der Künstlichen Intelligenz genutzt werden und diese damit gegen Angriffe absichern. (Vgl. Manoj Parmar & Amit Phadke, 2022)

AIShield führt eine Schwachstellenanalyse durch, die auf 200 identifizierten Angriffen auf KI-Modelle basiert. Nach dieser Schwachstellenanalyse des KI-Modells, wird eine Sicherheitsbarriere erzeugt, die eine automatische Verteidigung aus 14 validen Techniken

gewährleistet. Ein Schutzmechanismus den AIShield verwenden kann, um ein KI-System zu schützen ist, die Limitierung des Zugangs. Da ein Angreifer eine große Anzahl an Abfragen verwendet und sich dieser Angreifer so vom normalen Nutzer unterscheidet, kann dieser erkannt werden. Als Verteidigungsmaßnahme kann ein Schwellenwert an Abfragen definiert werden, der ausschließlich die festgelegte Anzahl an Abfragen zulässt. So kann das Entdecken von Schwachstellen für den Angreifer erschwert werden. (Vgl. Lekkala et al., 2021: S. 3; Tekwani & Parmar, 2022: S. 8)

Mithilfe verschiedenster Schutzmaßnahmen innerhalb von AIShield ist es möglich, die Genauigkeit eines stattfindenden Angriffs signifikant zu reduzieren, vergleiche Abbildung 29: Gegenüberstellung AIShield. In diesem Fall konnte die Genauigkeit eines Black-Box-Angriffs, mit einer großen Anzahl an Abfragen, durch die Verwendung von AIShield deutlich reduziert werden. Dabei ließ sich die Genauigkeit des Angreifers von 70% auf 15% reduzieren. Somit wird der Angriff enorm abgeschwächt, da der Angreifer weniger Nutzen aus dem Angriff ziehen kann. So wäre es durch die Verwendung von AIShield für das KI-Modell des Opfers, vergleiche Unterkapitel 6.3 Modell-Diebstahl, möglich gewesen, die Genauigkeit der Zeitreihenprognosen des Angreifers zu reduzieren. Damit hätte das Angreifer Modell eine deutlich schlechtere Genauigkeit im Vergleich zum Original. Bislang können Klassifikationen von Bildern geschützt werden. Jedoch ist AIShield noch nicht am Ende angelangt, da es weitere KI-Anwendungsgebiete gibt, die weiterhin ohne Schutz auskommen müssen, dazu gehören Zeitreihenprognosen und NLP. Deshalb ist es wichtig, Ansätze wie AIShield für die Zukunft weiter zu entwickeln, um diese Gebiete mit einem ausreichenden Schutz auszustatten.

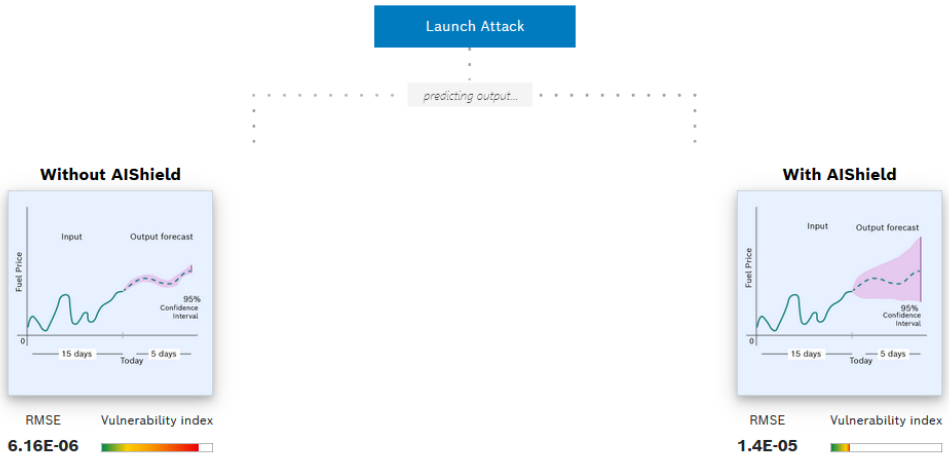

Abbildung 29: Gegenüberstellung AIShield (Quelle: Bosch AIShield, 2022b)

6.5 Zusammenfassung

Prognosen basierend auf der Vergangenheit durch Künstliche Intelligenz sind sehr wirkungsvoll und ermöglichen es, dadurch vorzeitig auf die Zukunft zu reagieren. Diese Zeitreihenprognosen lassen sich sowohl für Aufgaben im Büro als auch in der Fertigung verwenden und benötigen kein hochqualifiziertes Personal, das die Prognosen manuell durchführt. Die Künstliche Intelligenz kann diese Aufgabe übernehmen und bietet verschiedene KI-Modelle an, die dafür geeignet sind. Hierbei wurden neuronale Netzwerke verwendet, die in diesem Szenario, die Prognosen anhand von unterschiedlichen Features erstellen konnten. Im Weiteren war es möglich, die Funktionalität der erstellten KI-Modelle anhand der erstellten Prognosen zu stehlen. Diese Prognosen wurden anschließend verwendet, um ein Angreifer Modell zu trainieren. Die Performance der unterschiedlichen KI-Modelle kam der Perfomance vom Original bedeutend nahe. Ohne Schutz und Gegenmaßnahme ist es so möglich, ein aufwendiges und teures KI-Modell für Zeitreihenprognosen zu stehlen und innerhalb von kurzer Zeit zu trainieren. Deswegen sind Gegenmaßnahmen wie AIShield für die Zukunft von Künstlichere Intelligenz wichtig, um dagegen vorgehen zu können. Jedoch bieten die aktuellen Ansätze kaum Schutz oder ausschließlich für eine bestimmte KI-Variante. Wichtig für die Zukunft ist es, dass sich aktuelle Ansätze weiterentwickeln und neue entdeckt werden.

7 Zukunft der Cyber-Sicherheit für Künstliche Intelligenz

7.1 Kapitelübersicht

Die Bedeutung von Cyber-Sicherheit für Künstliche Intelligenz ist ein aktuelles und wichtiges Thema. Durch die immer weitreichendere Verwendung von Künstlicher Intelligenz, wird dieses Thema auch in Zukunft eine wahrscheinlich noch wichtigere Rolle als zum jetzigen Zeitpunkt einnehmen. Deshalb wird es von Bedeutung sein, nicht ausschließlich einzelne Ansätze zu finden, um dieses Problem anzugehen, sondern auch gemeinsame Ansätze, die von mehreren Parteien verfolgt und angewendet werden. Dabei werden Themen wie, Überwachung, Zertifizierung, oder auch die Reaktion auf Vorfälle innerhalb der Künstliche Intelligenz für die zukünftige Forschung von großer Bedeutung sein. Diese lassen sich weiterentwickeln oder anhand von ihnen andere Ansätze ableiten oder entdecken, die für Cyber-Sicherheit sorgen.

7.2 Systematisierung

Damit in Zukunft die Cyber-Sicherheit in Bezug auf Künstliche Intelligenz verstärkt werden kann, ist es nicht ausreichend, sich ausschließlich auf einen Aspekt zu konzentrieren. Deshalb ist es wichtig, dass verschiedene Faktoren zusammengeführt werden, um die Cyber-Sicherheit weiter zu steigern. Aus diesem Grund muss auch an Werkzeugen und Anwendungen wie AIShield weitergearbeitet werden, um mehr Anwendungsgebiete von Künstliche Intelligenz abzudecken und diesen Schutz zu gewährleisten, vergleiche Unterkapitel 6.4 Gegenmaßnahme AIShield. Wichtig für KI-Systeme ist es, dass diese überwacht werden und so Auffälligkeiten wahrgenommen werden können. Auf diese Weise können Angriffe aber auch abweichendes Verhalten erkannt werden und daraus weitere Schritte für die Absicherung getätigt werden.

Die Kennzeichnung von verschiedenen Angriffen unterstützt dabei, die unterschiedlichen Angriffe besser analysieren und kategorisieren zu können. Daraus lassen sich die verwendeten Schwachstellen der einzelnen Angriffe einfacher erkennen und Zusammenhänge erschlossen werden. Somit lässt sich die Reaktion auf einen Angriff anpassen und dadurch abgeschwächt oder ganz verhindert werden.

Um die Künstliche Intelligenz für Nutzer und Kunden vertrauenswürdiger gestalten zu können, wird eine Zertifizierung der KI-Systeme einen wichtigen Bestandteil für die Zukunft der Künstlichen Intelligenz darstellen, vergleiche Unterkapitel 7.4 Zertifizierung. Um ein Zertifikat zu erhalten, müssen bestimmte Anforderungen erfüllt sein und dies nicht nur für den Moment der Überprüfung, sondern auch zu einem späteren Zeitpunkt.

Angriffe können nicht immer gänzlich verhindert werden, deshalb muss bei einem laufenden Angriff schnell reagiert werden. Der Angriff muss zuerst erkannt werden, bevor die richtigen Maßnahmen getroffen werden können. Um das Ausmaß von aktiven Angriffen verringern zu können, müssen vorab Pläne definiert werden, die verwendet werden können. Auch die Bereinigung der Systeme gehört zu diesem Vorgang, sowie das Lernen aus diesen Situationen.

7.3 Überwachung

Künstliche Intelligenz bietet verschiedene Möglichkeiten und Ansätze, um sie zu überwachen und auf diese Weise eine effektive Kontrolle der KI-Anwendungen gewährleisten zu können. Eine Möglichkeit für eine Überwachung liegt in den Trainingsdaten. Dabei werden einkommende Daten, die als Trainingsdaten verwendet werden sollen überwacht und überprüft, ob sie eine gewisse Voreingenommenheit aufweisen oder absichtlich verändert wurden. Dieser Prozess wird in definierten Intervallen durchgeführt. Jedoch lässt sich die Anwendung selbst auch überwachen. Bei dieser Vorgehensweise werden die Ausgaben der KI-Anwendungen überwacht. Dafür müssen signifikante Merkmale oder auch potenziell gefährdete Gruppen während des gesamten Betriebs bekannt sein. Nur so können die KI-Anwendungen auf faire Ausgaben überprüft werden. Sowohl die Eingaben als auch die Ausgaben können dazu beitragen, dass die potenziellen Fehlerquellen ausgemacht werden können und anschließen behoben werden können. (Vgl. Maximilian Poretschkin et al., 2021: S. 46)

Nicht nur Änderungen an den Daten oder dem Modell sind wichtig zu überwachen, auch äußere Faktoren können sich ändern und müssen beachtet werden. Es können neue Arten von Diskriminierung erkannt werden, die noch nicht innerhalb der Trainingsdaten überprüft wurden. Auch ist es möglich, dass sich Änderungen in den Gesetzen durchsetzen und diese von einer KI-Anwendung eingehalten werden müssen. Möglichkeiten von Angriffen und Schadcode entwickeln sich schnell und müssen beachtet werden, deshalb ist es wichtig, auf dem neusten Stand der Forschung zu sein. Des Weiteren können Änderungen innerhalb der Software durch Updates zu Sicherheitslücken führen und sollten ausgiebig überwacht werden. Aus diesen Gründen ist es ratsam, die äußeren Faktoren zu überwachen und bei Änderungen eine passende Reaktion für die KI-Anwendungen einzuleiten. (Vgl. Maximilian Poretschkin et al., 2021: S. 140)

Die Überwachung von Künstlicher Intelligenz lässt sich durch unterschiedliche Ansätze umsetzen. Die Überwachung kann einerseits von einem Menschen durchgeführt werden oder durch einen kontinuierlichen automatischen Überwachungsprozess. Diese beide Methoden lassen sich zusammen verknüpfen und bieten somit eine Kombination dieser an. Bei diesem

Ansatz lassen sich die KI-Anwendungen automatisch überwachen und erfassen währenddessen Veränderungen von Eingabedaten im Betrieb. Die erhaltenen Informationen werden ausgewertet und bei der Feststellung von kritischen oder neuartigen Eingaben, werden diese dauerhaft abgespeichert. Diese Eingaben werden für weitere Trainingseinheiten und zur Verbesserung der Detektion verwendet, umso die Verlässlichkeit zu erhöhen. Sollten jedoch zu große Änderungen anfallen oder die Verlässlichkeit nicht mehr gewährleistet sein, so wird dem Menschen mitgeteilt, dass eine Aktualisierung oder die Abschaltung des KI-Systems durchgeführt werden muss, damit die Sicherheit gewährleistet werden kann. (Vgl. Maximilian Poretschkin et al., 2021: S. 114)

7.4 Zertifizierung

In verschiedenen Lebensbereichen konnten bereits Standards und Richtlinien implementiert werden, die an Bedeutung zugelegt haben. Alleine sind diese jedoch unverbindlich. Durch die zunehmende Bedeutung von Standards und Richtlinien werden diese entweder vom Gesetzgeber verpflichtend gemacht oder von Auftraggebern verlangt, dass die enthaltenen Normen eingehalten werden. Die Überprüfung der Einhaltung kann selbst oder durch Partner vorgenommen werden. Um das größte Maß an Konformität zu den Normen gewährleisten zu können, gibt es die Möglichkeit, sich durch eine dritte Instanz zertifizieren zu lassen, die über die notwendige Kompetenz verfügt. So kann gewährleistet werden, dass nach dem Stand von Technik und Wissenschaft gehandelt wurde. (Vgl. Axel Mangelsdorf et al., 2021: S. 3)

Damit Schwachstellen innerhalb der Künstlichen Intelligenz in Zukunft besser erkannt werden können und um die Angriffe zu vergleichen, ist es wichtig, ein Verständnis der Angriffe und der Angriffsfläche zu entwickeln. Dies trägt dazu bei, dass durch die Kennzeichnung, die Zertifizierung der KI-Systeme vereinfacht wird. Es gibt immer mehr Angriffe, die Auswirkungen auf KI-Systeme haben, vergleiche Unterkapitel 4.2 Adversarial Künstliche Intelligenz. Deshalb besteht der Bedarf, ein gemeinsames Verständnis zu entwickeln und dieses mithilfe von Kennzeichnungen zu unterstützen und erweitern. Angriffe auf KI-Systeme haben unterschiedliche Gründe und Ziele, deshalb ist es beutend, diese zu kennzeichnen und einzustufen. Dies ermöglicht Sicherheitsrisiken innerhalb der Entwicklungsphase von KI-Systemen zu erkennen und angehen zu können. Dabei lässt sich der Zusammenhang zwischen Risiko und Schwachstelle besser herausfinden und somit auch das Potential eines Angriffs. Durch die weitere Integrierung von KI-Systemen in der nahen Zukunft, wird es an Bedeutung zunehmen, die Schwachstellen von den KI-Systemen zu identifizieren und zu verstehen. (Vgl. Hartmann & Steup, 2020: S. 335 f.)

Durch eine Einstufung von Angriffen ist es möglich, verschiedene Angriffe zusammenzufassen und die dahinterliegenden Ziele zu erkennen. Neue Angriffe können zu einer bereits identifizierten Einstufung einer Angriffsart hinzugefügt werden. Daraus können sich effizienter Vorkehrungen gegen dieselbe Art von Angriff gestalten lassen, als einen Ansatz für jeden einzelnen Angriff. Die Einstufung von Angriffen in Bezug auf KI-Modelle kann sinnvoll sein, da die KI-Modelle sich in ihrer Funktionsweise unterscheiden und somit mehr oder weniger Einstiegspunkte für Angreifer bieten können. Wichtig dabei ist zu beachten, dass Schwachstellen für Angreifer durch KI-Architekturen, Methoden, Implementierungsentscheidungen aber auch durch die Speicherung und Verarbeitung der Daten entstehen können. Dies muss bei Kennzeichnung berücksichtigt werden, um Kenntnisse gegenüber Schwachstellen und Angriffen von KI-Systemen zu schaffen. (Vgl Hartmann & Steup, 2020: S. 345-347)

KI-Systeme benötigen ein hohes Maß an Vertrauen, dass diese sicher, fair und zuverlässig sind. Deshalb ist es eine Schlüsselvoraussetzung, die KI-Anwendungen durch unabhängige Dritte zertifizieren zu lassen, umso das Vertrauen in die Künstliche Intelligenz zu steigern und auch die Sicherheit und Fairness dieser zu verbessern. Die Potentiale der Zertifizierung im Bereich Künstliche Intelligenz liegen darin, dass sie die Vertrauenswürdigkeit der KI-Systeme steigern aber auch dieses Vertrauen auf die Hersteller und Anbieter übertragen. Die Zertifizierung ist nicht dafür gedacht, alle mögliche negativen Konsequenzen ausschließen zu können. Vielmehr kann mit der Zertifizierung ein Großteil von ihnen verhindert werden und stellt somit eine Minimalanforderung dar. Allerdings muss dabei beachtet werden, dass Künstliche Intelligenz ein weitgefächertes Spektrum an Anwendungsgebieten besitzt und somit die Effizienz der Zertifizierung in diesem Themenbereich weiter erforscht werden muss. (Vgl. Axel Mangelsdorf et al., 2021: S. 2; Jessica Heesen et al., 2020: S. 3)

Eine Zertifizierung von KI-Systemen ist notwendig, da diese im Gegensatz zu den herkömmlichen IT-Systemen dynamisch sind. Sie entwickeln sich weiter und dies kann auch unbeabsichtigt passieren. Die Entscheidungen, die von einer KI-Anwendung getroffen werden, können fehlerbehaftet sein und sind außerdem schwer nachzuvollziehen. Auch wird in der Zukunft die Interaktion zwischen Mensch und Maschine weiter zunehmen. Die Zertifizierung kann dazu beitragen, dass bei entstehenden Schäden, die Verantwortungs- und Haftungsfragen geregelt sind. Durch die Zertifizierung wird bestätigt, dass gesellschaftliche und ökonomische Kriterien erfüllt werden. Dabei stehen Prinzipien wie Rechtssicherheit, IT-Sicherheit, Datenschutz aber auch Transparenz und Verantwortlichkeit im Fokus und auch viele weitere Prinzipien. (Vgl. Jessica Heesen et al., 2020: S. 4-6)

Durch die Zertifizierung wird eine Vergleichbarkeit zwischen den KI-Systemen für die Gesellschaft geboten. Durch die Bestätigung der Einhaltung der Kriterien und Prinzipien, können die Akteure aus diesem Bereich, die unterschiedlichen KI-systeme selbst für sich Beurteilen und eine eigene Wahl treffen. Dadurch ist es möglich, dass die Verständlichkeit und die Akzeptanz von Künstlicher Intelligenz steigen. Die festgelegten Kriterien unterstützen bei dem Entwicklungsprozess, indem sich Entwickler daran orientieren können und so die definierten Kriterien von Anfang an mitverwenden. Auf diese Weise kann aus Herstellerperspektive mehr Absatz für KI-Produkte generiert und den Wert dieser gesteigert werden. (Vgl. Jessica Heesen et al., 2020: S. 6)

Jedoch stehen der Zertifizierung für Künstliche Intelligenz Herausforderungen gegenüber, die angegangen werden müssen. Durch das falsche Maß an Zertifizierung kann der Markt negativ beeinflusst werden, indem potenzielle Kosten für Unternehmen steigen oder den Markt an sich verwehren. Des Weiteren fehlen Erfahrungswerte, um die Dynamik von Künstlicher Intelligenz zu zertifizieren. Das Verhalten von KI-Systemen ist nicht immer vorhersehbar und die Umwelt kann sich schlagartig verändern. Deshalb ist es wichtig, nicht ausschließlich eine Momentaufnahme zu zertifizieren. Vielmehr sollte eine kontinuierliche Zertifizierung stattfinden. (Vgl. Axel Mangelsdorf et al., 2021: S. 6 f.; Jessica Heesen et al., 2020: S. 10 f.)

7.5 Reaktion auf KI-Vorfälle

KI-Vorfälle können nicht immer verhindert werden, deshalb ist es essenziell, auf diese Situationen vorbereitet zu sein, vergleiche Kapitel 5 Die Bedeutung von Cyber-Sicherheit Governance in Bezug auf Künstliche Intelligenz. So kann schnell auf die vorliegende Situation reagiert werden. Zunächst muss jedoch ein KI-Vorfall erkannt werden und dies kann bei KI-Systemen eine große Herausforderung darstellen. Danach kann der Prozess eines Notfall-Managements, der zuvor geplant wurde, durchgeführt werden. Nachdem ein Fehler oder ein Gefährdungsszenario erkannt wurde, muss die KI-Anwendung Sicherheit gewährleisten. Dabei ist es abhängig, welche Auswirkungen durch einen Angriff oder Fehler entstehen. Dabei kann das KI-System entweder in einen Überbrückungszustand überführt werden, indem die Funktionalität aufrechterhalten wird oder die Abschaltung eingeleitet werden, um in einen Fail-Safe-Modus überzugehen. Dabei wird durch den sicheren Zustand versucht, den potenziellen Schaden so gering wie möglich zu halten. So kann bei bekannt werden der Gewichte eines KI-Modells, das Risiko eines White-Box-Angriffs, beispielswiese durch Neutraining gemindert werden. (Vgl. Maximilian Poretschkin et al., 2021: S. 137)

Die enthaltenen Notfallkonzepte innerhalb des Notfall-Managements, sollten nicht nur geplant und implementiert werden, sondern auch getestet und verbessert, um in Notfallsituationen vorbereitet zu sein. Bei einem Ernstfall ist es wichtig, eine Notfall-Analyse durchzuführen und die Daten sowie das KI-Modell zu überprüfen. In der Notfall-Analyse werden die eingetretenen Notfälle protokolliert, um sie auswerten und analysieren zu können. Aus den erlangten Informationen können die implementierten Sicherheitsmaßnahmen darauf überprüft werden, wie wirksam diese waren. Auf Basis der erlangten Erkenntnisse können diese angepasst oder ausgetauscht werden. Des Weiteren lässt sich dadurch die Sicherheit der Anwendung weiter verbessern. Die Adversarial Daten können für einen neuen Trainingsprozess verwendet werden, um die Robustheit des KI-Systems gegen die Angreifer Daten zu erhöhen. (Vgl. Maximilian Poretschkin et al., 2021: S. 140; Qiu et al., 2019: S. 20 f.)

Damit die Daten, die durch äußere Einwirkungen beeinflusst werden, wiederverwendet werden können, sind Maßnahmen wie zum Beispiel Backups essenziell. Um ein angegriffenes KI-Modell von einem Angriff zu bereinigen, muss es möglich sein, das Modell auf die letzte Version zurückzusetzen. Darum sind Backups für KI-Modelle wichtig, um diese auf demselben Niveau zu halten. Bei Backups ist zu berücksichtigen, dass diese regelmäßig gespeichert und auf Funktion sowie Änderungen überprüft werden. Sollten unerwünschte Änderungen von einer KI-Komponente nicht zu beheben sein, so muss diese neu aufgesetzt werden. (Vgl. Maximilian Poretschkin et al., 2021: S. 136 f.)

7.6 Zusammenfassung

Damit Künstliche Intelligenz durch Cyber-Sicherheit in Zukunft besser geschützt werden kann, ist es essenziell, dass diese weiter ausgebaut und verbessert wird. Eine Maßnahme dafür ist die Überwachung, diese kann auf die Trainingsdaten bezogen werden oder auf die KI-Anwendung selbst, indem die Ausgaben überwacht werden. Diese lassen sich entweder durch den Menschen oder durch andere KI-Anwendungen überwachen. Des Weiteren kann eine Zertifizierung dabei helfen, das Vertrauen und die Fairness für Künstliche Intelligenz zu verbessern. Einem Verwender wird dadurch signalisiert, dass ein gewisses Maß an Sicherheit vorherrscht und kann dies so mit anderen Anbietern vergleichen. Auch wird dadurch ermöglicht, ein gemeinsames Verständnis für Gefahren zu entwickeln, das dazu beitragen kann, unterschiedliche Angriffsarten besser einstufen zu können. Der Zertifizierung steht jedoch das Problem gegenüber, dass sich KI-Systeme und die Umwelt rasant ändern können und so die Zertifizierung zum aktuellen Stand nicht gewährleistet werden kann. Um einen aktiven Schutz gegen Angriffe auf KI-Systeme zu besitzen, sind Adversarial Trainings wichtig, um die KI-

Systeme darauf vorzubereiten, angegriffen zu werden. Ein Notfall-Management kann bei einem aktiven Angriff davor schützen, dass größer Schäden entstehen. Dabei wird das KI-System in einen sichern Zustand überführt. Für eine Wiederherstellung der Daten nach einem erfolgreichen Angriff sind Backups ein wesentlicher Bestandteil der Vorbereitung für den Ernstfall.

8 Fazit

8.1 Zusammenfassung

Künstliche Intelligenz entwickelt sich weiter und erzielt immer bessere Ergebnisse. Es ergeben sich neue Anwendungsgebiete, in denen sich die Künstliche Intelligenz implementieren lässt und so Schritt für Schritt mehr Einfluss auf unser Leben nehmen wird. Für den Erfolg der Künstlichen Intelligenz spielen unterschiedliche Faktoren eine wichtige Rolle. Dazu gehören die unterschiedlichen Modelle, die je nach Anwendungsgebiet eine bessere oder schlechtere Performance leisten können. Zeitreihenprognosen besitzen eine große Bedeutung für Unternehmen und durch die Verwendung von Künstliche Intelligenz in diesem Bereich, wird diese Bedeutung immer wichtiger.

Die Informationstechnologie stellt ein beliebtes Angriffsziel dar. Dabei gibt es verschiedenste Absichten und Ziele, die verfolgt werden, um entweder Schaden anzurichten oder sich selbst einen Vorteil zu verschaffen. Dafür stehen diverse Angriffsmöglichkeiten zur Verfügung, die unterschiedlichste Schwachstellen ausnutzen. Dazu gehört auch die Schwachstelle des Menschen. Jedoch sind für die meisten Angriffsmöglichkeiten Gegenmaßnahmen bekannt, um die Gefahr eines erfolgreichen Angriffs zu reduzieren oder ganz zu verhindern. Dennoch sind einzelne Gegenmaßnahmen nur der Anfang für einen guten Schutz für die Informationstechnologie. Damit ein besserer und weit gefächerter Schutz gewährleistet werden kann, ist IT-Governance in Bezug auf Cyber-Sicherheit ein essenzieller Bestandteil. Cyber-Sicherheit mit IT-Governance verknüpft, stellt ein ganzheitliches Konzept dar, dass das Sicherheitsniveau für alle Geschäftsprozesse, Daten als auch IT-Systemen in Organisationen etabliert. Der dahinter liegende Prozess wird kontinuierlich überprüft und bei Bedarf werden Anpassungen vorgenommen, damit eine ständige Weiterentwicklung und somit auch Verbesserungen stattfinden können. Nur so ist es möglich, gegen neuste Angriffsmethoden gewappnet und auf dem neusten Sicherheitsstand zu sein.

Künstliche Intelligenz kann nicht nur für gute Zwecke verwendet werden, auch ist es möglich, diese anzugreifen oder mit böswilligen Absichten zu verwenden. So ist es möglich, durch Angriffe auf die KI-Systeme, immense Schäden anzurichten oder Lebewesen zu gefährden. Dies spiegelt sich bei Zeitreihenprognosen wider, da hierbei Schäden durch falsche Kalkulation oder Entwendung entstehen. Je nach Angriffsszenario werden die Daten für das Training oder für die Vorhersagen manipuliert, um daraus falsche Prognosen zu generieren. Oder es wird versucht, die Funktionalität zu entwenden, um diese für eigene Zwecke zu verwenden. Auch können Angreifer eigene KI-Modelle erschaffen, die Angriffe effizienter gestalten und die

Wahrscheinlichkeit erfolgreich zu sein erhöhen. Aus diesem Grund ist es wichtig, die Künstliche Intelligenz für Cyber-Sicherheit zu verwenden. Dadurch lassen sich Angriffe schneller erkennen und Kategorisieren. Dadurch wird es möglich, effizienter auf unterschiedliche Angriffe zu reagieren aber auch neue Angriffe zu detektieren. Für eine Gewährleistung von sicherer Künstlicher Intelligenz und Schutz durch Künstliche Intelligenz, müssen Vorkehrungen zur Absicherung getroffen werden. So kann das Risiko, das im Bereich Künstlicher Intelligenz besteht, gemanagt werden. Jedoch muss der gesamte KI-Lebenszyklus überwacht werden, um die Sicherheit durch diesen KI-Governance-Prozess in jeder Lage beurteilen zu können.

Für eine vollumfängliche Cyber-Sicherheit für Künstliche Intelligenz müssen die zu schützenden Wertgegenstände bekannt sein. Durch dieses Wissen kann das jeweilige Risiko eingeschätzt und Sicherheitsmaßnahmen abgeleitet werden, die das zugrundeliegende Risiko so weit minimieren, dass es akzeptabel ist. Die Wertgegenstände, die für Künstliche Intelligenz auf jeden Fall zu schützen sind, sind die verwendeten Daten aber auch die Modelle und ihren dazugehörigen Parametern. Zufällige Änderungen durch Fehler oder Angreifer können zu großen Auswirkungen führen. Diese Auswirkungen können sowohl im technischen als auch im sozio-technischen Bereich liegen und somit unterschiedlichen Einfluss ausüben. Um Fehler und Schwachstellen zu erkennen und somit die daraus resultierenden Auswirkungen anzugehen, müssen Schutzvorkehrungen für KI-Systeme getroffen werden. Für Künstliche Intelligenz ist es wichtig, immer auf dem neusten Stand der Forschung bezüglich neuer Angriffe zu sein. Des Weiteren ist es wichtig, Zugangsberechtigungen zu erteilen, um Änderungen auszuschließen und dass andere herkömmlichen Schutzvorkehrungen durchgeführt werden. Um aktiv Schwachstellen zu erkennen, sollten die KI-Systeme von Experten gezielt angegriffen werden. Um spezifisch gegen Adversarial Angriffe aktiv vorzugehen, müssen die KI-Systeme durch Adversarial Training gepflegt oder aktiv auf Anomalien geprüft werden.

Für die Fallstudie wurden unterschiedliche KI-Modelle für Zeitreihenprognosen verwendet, um diese dem Angriffsszenario Modell-Diebstahl auszusetzen. Dabei wurden normalisierte Daten verwendet, die den Angriff unterstützt haben. Diese wurden vom originalen Modell verwendet und anhand von diesen Prognosen erzeugt, die für ein eigenes Angreifer Modell verwendet werden konnten. Gegenmaßnahmen wie AIShield können dazu beitragen, dass KI-Modelle geschützt und solche Vorfälle verhindert werden können.

Deshalb ist es wichtig, so früh wie möglich die Cyber-Sicherheit in Bezug auf Künstliche Intelligenz zu verbessern. Um mit der rasanten Entwicklung mithalten zu können, ist es wichtig,

Überwachung, Zertifizierung, Kennzeichnung und die Reaktion auf KI-Vorfälle mehr zu beachten und so für die Zukunft zu verbessern.

8.2 Ergebnisse

In dieser Arbeit konnte festgestellt werden, dass Cyber-Sicherheit für Künstliche Intelligenz weitestgehend am Anfang steht. Es gibt wenige Ansätze, die sich mit der Absicherung von KI-Anwendungen auseinandersetzen. Bereits in der herkömmlichen IT von vielen Unternehmen ist Cyber-Sicherheit nicht vollumfänglich angegangen worden. Dies führt dazu, dass sich das Bewusstsein für die Cyber-Sicherheit für Künstliche Intelligenz langsam entwickelt. **Jedoch wächst die ausgehende Gefahr gegenüber Künstlicher Intelligenz schneller als die Gegenmaßnahmen**. Deshalb ist es von großer Bedeutung, das ausgehende Risiko der unsicheren KI-Systeme so früh wie möglich zu reduzieren. Dann ist es möglich, das Vertrauen gegenüber Künstlicher Intelligenz zu stärken und das volle Potential dieser auszunutzen.

Dazu gehört es nicht nur eine Absicherung gegenüber Angriffen zu implementieren. Es muss schon bei der Erstellung von KI-Systemen darauf geachtet werden, dass diese fair, transparent und präzise sind. Ansonsten kann das ausgehende Risiko nicht soweit reduziert werden, dass es ein akzeptables Niveau erreicht. Jeder Wertgegenstand einer Künstlichen Intelligenz kann einen Risikofaktor darstellen und muss deshalb durch einen kontinuierlichen Prozess abgesichert werden. Es ist wichtig, dass dieser Prozess den neusten Stand der Forschung verwendet und immer weiterentwickelt wird.

Die Verwendung von Künstlicher Intelligenz benötigt viel Zeit, Aufwand und finanzielle Mittel. Dies ist unabhängig von dem Anwendungsbereich, ob Bilderklassifizierung, Zeitreihenprognosen oder in einem anderen Bereich. Damit die investierte Arbeit nicht zerstört oder entwendet werden kann, sind die aktuell wenigen vorhandenen Gegenmaßnahmen ein guter Anfang. Jedoch müssen diese weiterentwickelt und neue Ansätze dafür geschaffen werden. Dies ist nur möglich, wenn die Gefahren gegenüber Künstliche Intelligenz mehr in das Bewusstsein integriert werden.

Essenziell für die Zukunft für Künstliche Intelligenz wird die Cyber-Sicherheit sein. Durch Zertifizierungen als Sicherheitsmechanismus für Künstliche Intelligenz, kann die Sicherheit erhöht und gleichzeitig die Vermarktung der KI-Produkte erfolgreicher werden. Dafür werden einheitliche Standardisierungen notwendig sein, die eingehalten werden müssen oder zur Orientierung dienen.

8.3 Ausblick

Künstliche Intelligenz wird sich immer weiter ausbreiten und sich in unser Leben integrieren. Diese Auswirkung werden wir nicht nur in neuen Anwendungsbereichen bemerken, sondern auch durch die Weiterentwicklung der Künstlichen Intelligenz und der daraus folgenden steigenden Effizienz. Damit dies ohne Gefahren geschehen kann, muss die Cyber-Sicherheit für Künstliche Intelligenz in vielen Bereichen aufholen. Denn durch eine sichere Künstliche Intelligenz, steigt das Vertrauen der Menschen in diese sowie das Potential, das die KI mit sich bringt. Dafür ist es jedoch notwendig, das Bewusstsein für potenzielle Gefahren weiter zu schärfen und die bisherigen Ansätze weiter auszubauen sowie neue Ansätze für den Schutz von KI-Systemen zu finden. Für den Anfang wäre es sinnvoll, eine einheitliche Basis zu schaffen, die ein gewisses Maß an Grundsicherheit bietet. So kann ein Minimum an Sicherheit gewährleistet werden und darauf aufbauend, weitere Maßnahmen und Entwicklungen für die Sicherheit der Künstlichen Intelligenz integriert werden.

Literaturverzeichnis

Adi Ashkenazy & Shahar Zini (2019) *Bypassing Cylance's AI Malware Detection, Case Study: AML.CS0003 | MITRE ATLAS™* [Online]. Verfügbar unter https://atlas.mitre.org/studies/AML.CS0003 (Abgerufen am 13 Mai 2022).

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A. & Almomani, E. (2013) „A Survey of Phishing Email Filtering Techniques“, *IEEE Communications Surveys & Tutorials*, Vol. 15, No. 4, S. 2070–2090.

Alrajeh, N. A. & Lloret, J. (2013) „Intrusion Detection Systems Based on Artificial Intelligence Techniques in Wireless Sensor Networks“, *International Journal of Distributed Sensor Networks*, Vol. 9, No. 10, S. 351047.

Aslan, O. & Samet, R. (2020) „A Comprehensive Review on Malware Detection Approaches“, *IEEE Access*, Vol. 8, S. 6249–6271.

Axel Mangelsdorf, Peter Gabriel & Martin Weimer (2021) *Die Zertifizierung von KI: Mehr Sicherheit für alle–oder unnötiger Ballast?* [Online], 58. Aufl., Berlin, Institut für Innovation und Technik (iit). Verfügbar unter https://www.iit-berlin.de/wp-content/uploads/2021/04/2021_04_30_iit-perspektive_Nr-58_Zertifizierung_von_KI.pdf (Abgerufen am 18 August 2022).

Bauer, L. (2021) „Digital-Trend-Studie-Kuenstliche-Intelligenz-data“ [Online]. Verfügbar unter https://www.dbsystel.de/resource/blob/6075020/d8e52123113be2057ba86ba06eaddedc/Digital-Trend-Studie-Kuenstliche-Intelligenz-data.pdf (Abgerufen am 16 Juni 2022).

Beitollahi, H. & Deconinck, G. (2012) „Analyzing well-known countermeasures against distributed denial of service attacks“, *Computer Communications*, Vol. 35, No. 11, S. 1312–1332.

Bhatti, B. M., Mubarak, S. & Nagalingam, S. (2021) „Information security implications of using NLP in IT outsourcing: a Diffusion of Innovation theory perspective“, *Automated Software Engineering*, Vol. 28, No. 2, S. 1–29 [Online]. DOI: 10.1007/s10515-021-00286-x (Abgerufen am 18 August 2022).

Bonfanti, M. E. & Kohler, K. (2020) *Künstliche Intelligenz für die Cybersicherheit* [Online]. Verfügbar unter https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/417506/CSSAnalyse265-DE.pdf?sequence=2&isAllowed=y (Abgerufen am 29 April 2022).

Bosch AIShield (2022a) [Online]. Verfügbar unter https://www.boschaishield.com/ (Abgerufen am 29 Juli 2022).

Bosch AIShield (2022b) [Online]. Verfügbar unter https://aishieldwebdemo.z13.web.core.windows.net/#sectionone (Abgerufen am 17 Juli 2022).

Bostrom, N. „Superintelligence: Paths, Dangers, Strategies“ [Online]. Verfügbar unter https://dorshon.com/wp-content/uploads/2017/05/superintelligence-paths-dangers-strategies-by-nick-bostrom.pdf (Abgerufen am 29 Mai 2022).

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2016) *Time series analysis: Forecasting and control*, 5. Aufl., Hoboken, Wiley.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R. & Amodei, D. (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* [Online]. Verfügbar unter https://arxiv.org/pdf/1802.07228 (Abgerufen am 18 August 2022).

Buber, E., Diri, B. & Sahingoz, O. K. (2018) „NLP Based Phishing Attack Detection from URLs“, in Abraham, A., Muhuri, P. K., Muda, A. K. & Gandhi, N. (Hg.) *INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS: 17th international conference on*, SPRINGER INTERNATIONAL PU, S. 608–618.

Buchanan, B. (2020) *A National Security Research Agenda for Cybersecurity and Artificial Intelligence*, Center for Security and Emerging Technology DOI: 10.51593/2020CA001 (Abgerufen am 18 August 2022).

Bundesamt für Sicherheit in der Informationstechnik (2017) „BSI-Standard 200.2“ [Online]. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/BSI_Standards/standard_200_2.pdf?__blob=publicationFile&v=2 (Abgerufen am 18 Juni 2022).

Bundesamt für Sicherheit in der Informationstechnik (2021) „Sicherer, robuster und nachvollziehbarer Einsatz von KI - Probleme, Maßnahmen und Handlungsbedarfe“ [Online]. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=6 (Abgerufen am 24 April 2022).

Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik, Verband der TÜV e.V. (2021) „Whitepaper: Towards Auditable AI Systems“ [Online]. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=4 (Abgerufen am 24 April 2022).

Cambria, E. & White, B. (2014) „Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]“, *IEEE Computational Intelligence Magazine*, Vol. 9, No. 2, S. 48–57.

Candanedo, L. M., Feldheim, V. & Deramaix, D. (2017) „Data driven prediction models of energy use of appliances in a low-energy house“, *Energy and Buildings*, Vol. 140, S. 81–97 [Online]. DOI: 10.1016/j.enbuild.2017.01.083 (Abgerufen am 18 August 2022).

Chandola, V., Banerjee, A. & Kumar, V. (2009) „Anomaly detection“, *ACM Computing Surveys*, Vol. 41, No. 3, S. 1–58 [Online]. DOI: 10.1145/1541880.1541882 (Abgerufen am 18 August 2022).

Christiaan Beek (2020) *VirusTotal Poisoning, Case Study: AML.CS0002 | MITRE ATLAS™* [Online]. Verfügbar unter https://atlas.mitre.org/studies/AML.CS0002 (Abgerufen am 13 Mai 2022).

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. & Bharath, A. A. (2018) „Generative Adversarial Networks: An Overview“, *IEEE Signal Processing Magazine*, Vol. 35, No. 1, S. 53–65.

Deutscher Dialogmarketing Verband e. V., D. D. (2019) *Dialogmarketing Perspektiven 2018/2019: Tagungsband 13. Wissenschaftlicher Interdisziplinärer Kongress Für Dialogmarketing*, Wiesbaden, Gabler.

Deutsches Institut für Normung e. V (2015) *27001:2015-03: Informationstechnik – IT-Sicherheitsverfahren – Informationssicherheits-Managementsysteme – Anforderungen*, Berlin: Beuth Verlag GmbH.

Dirk Hecker, Inga Döbel, Ulrike Petersen, André Rauschert, Velina Schmitz & Angelika Voss (2018) *Zukunftsmarkt Künstliche Intelligenz–Potenziale und Anwendungen* [Online], Frauenhofer-Allianz. Verfügbar unter https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/KI-Potenzialanalyse_2017.pdf (Abgerufen am 18 August 2022).

Dolado, J. J., Gonzalo, J. & Mayoral, L. (2002) „A Fractional Dickey-Fuller Test for Unit Roots“, *Econometrica*, Vol. 70, No. 5, S. 1963–2006.

European Commission (2017) *Lage der Union 2017 – Cybersicherheit: Kommission will Reaktionsfähigkeit der EU auf Cyberangriffe verbessern* [Online]. Verfügbar unter https://ec.europa.eu/commission/presscorner/detail/de/IP_17_3193 (Abgerufen am 3 August 2022).

European Court of Auditors (2019) „Challenges to effective EU cybersecurity policy“ [Online]. Verfügbar unter https://www.eca.europa.eu/Lists/ECADocuments/BRP_CYBERSECURITY/BRP_CYBERSECURITY_DE.pdf (Abgerufen am 4 Juni 2022).

European Union Agency for Cybersecurity (2020) *Artificial Intelligence Cybersecurity Challenges; Threat Landscape for Artificial Intelligence* [Online]. Verfügbar unter https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges (Abgerufen am 29 Juli 2022).

Europol (2018) *World’s biggest marketplace selling internet paralysing DDoS attacks taken down | Europol* [Online], Europol. Verfügbar unter https://www.europol.europa.eu/media-press/newsroom/news/world%e2%80%99s-biggest-marketplace-selling-internet-paralysing-ddos-attacks-taken-down#downloads (Abgerufen am 6 Juni 2022).

Executive Office of the President (2016) „PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE“ [Online]. Verfügbar unter https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (Abgerufen am 30 Mai 2022).

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. & Song, D. (2017) *Robust Physical-World Attacks on Deep Learning Models* [Online]. Verfügbar unter https://arxiv.org/pdf/1707.08945 (Abgerufen am 18 August 2022).

Fang, J., Su, H. & Xiao, Y. (2018) „Will Artificial Intelligence Surpass Human Intelligence?“, *SSRN Electronic Journal* [Online]. DOI: 10.2139/ssrn.3173876 (Abgerufen am 27 Mai 2022).

Federal Office for Information Security (2021) „AI Cloud Service Compliance Criteria Catalogue“ [Online]. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf;jsessionid=CE7373AB1039E245C31F38930D7D7B0F.internet471?__blob=publicationFile&v=4 (Abgerufen am 24 April 2022).

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2014) *Explaining and Harnessing Adversarial Examples* [Online]. Verfügbar unter https://arxiv.org/pdf/1412.6572.pdf (Abgerufen am 1 Juni 2022).

Haluk Demirkan, Seth Earley & Robert R. Harmon (2017) „Cognitive Computing“ [Online]. Verfügbar unter https://ieeexplore.ieee.org/ielx7/6294/8012274/08012289.pdf?tp=&arnumber=8012289&isnumber=8012274&ref= (Abgerufen am 1 Juni 2022).

Hans P. Reiser, Noëlle Rakotondravony & Johannes Köstler (2017) *Mikromodul 8002: Cloud-Sicherheit und Bedro-hungsmodelle* [Online]. Verfügbar unter https://www.fim.uni-passau.de/fileadmin/dokumente/fakultaeten/fim/lehrstuhl/reiser/openc3s/CloudSecFor-MM-8002.pdf (Abgerufen am 17 August 2022).

Happiness Ugochi Dike, Yimin Zhou, Kranthi Kumar Deveerasetty & Qingtian Wu (2018) *Unsupervised Learning Based On Artificial Neural Network: A Review* [Online]. Verfügbar unter https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8612259&tag=1 (Abgerufen am 21 Juni 2022).

Hartmann, K. & Steup, C. (2020) „Hacking the AI - the Next Generation of Hijacked Systems“, *2020 12th International Conference on Cyber Conflict (CyCon).* Estonia, 5/26/2020 - 5/29/2020. [S.l.], IEEE, S. 327–349.

Hauschke Andreas & Hildebrandt Stefanie (2022) „vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data“ [Online]. Verfügbar unter https://www.vde.com/resource/blob/2176686/a24b13db01773747e6b7bba4ce20ea60/vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data.pdf (Abgerufen am 3 Juli 2022).

Hochreiter, S. & Schmidhuber, J. (1997) „Long short-term memory“, *Neural Computation*, Vol. 9, No. 8, S. 1735–1780.

Hölldobler, S & Gesellschaft für Informatik e.V., BB (Hg.) (2021) *GI LNI Dissertations Band 20 - Ausgezeichnete Informatikdissertationen 2019*, Bonn, Köllen.

Huseyin Cavusoglu, Hasan Cavusoglu & Jun Zhang (2006) *Economics of Security Patch Management* [Online]. Verfügbar unter https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.767&rep=rep1&type=pdf (Abgerufen am 18 August 2022).

Imran, A. A., Amin, M. N., Islam Rifat, M. R. & Mehreen, S. (2019 - 2019) „Deep Neural Network Approach for Predicting the Productivity of Garment Employees“, *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT).* Paris, France, 4/23/2019 - 4/26/2019, IEEE, S. 1402–1407.

Internet Organised Crime Threat Assessment (IOCTA) 2018 | Europol (2022) [Online]. Verfügbar unter https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2018 (Abgerufen am 15 Juni 2022).

Jan-Hendrik Meier, Stephan Schneider, Holm Voss & Anna-Katharina Dhunge (2021) *DIGITALE VERNETZUNG, SELBSTSCHÜTZENDE VERNETZUNG, DATENANALYSE UND DATA MINING* [Online]. Verfügbar unter https://d-nb.info/1231253711/34#page=175 (Abgerufen am 12 Juli 2022).

Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel, Maximilian Poretschkin, Stephanie Dachsberger & Maximilian Hösl (2020) *Zertifizierung von KI-Systemen – Impulspapier aus der Plattform Lernende Systeme* [Online], München. Verfügbar unter https://www.ml2r.de/wp-content/uploads/PDFs/AG3_Impulspapier_290420.pdf (Abgerufen am 18 August 2022).

Juuti, M., Szyller, S., Marchal, S. & Asokan, N. (2018) *PRADA: Protecting against DNN Model Stealing Attacks* [Online]. Verfügbar unter https://arxiv.org/pdf/1805.02628.pdf (Abgerufen am 24 April 2022).

Karbab, E. B. & Debbabi, M. (2019) „MalDy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports“, *Digital Investigation*, Vol. 28, S77-S87 [Online]. DOI: 10.1016/j.diin.2019.01.017 (Abgerufen am 18 August 2022).

Karunasingha, D. S. K. (2022) „Root mean square error or mean absolute error? Use their ratio as well“, *Information Sciences*, Vol. 585, S. 609–629 [Online]. DOI: 10.1016/j.ins.2021.11.036 (Abgerufen am 18 August 2022).

Kumar, N. & Susan, S. (2020?) „COVID-19 Pandemic Prediction using Time Series Forecasting Models“, *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Kharagpur, India, 7/1/2020 - 7/3/2020. [Piscataway, New Jersey], IEEE, S. 1–7.

L. N. Long & C. F. Cotner (2019) „A Review and Proposed Framework for Artificial General Intelligence“, *2019 IEEE Aerospace Conference*, S. 1–10.

Laura Pullum (2022) *Verification and Validation of Systems in Which AI is a Key Element - SEBoK* [Online]. Verfügbar unter https://www.sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element (Abgerufen am 31 Juli 2022).

Laurent Dupont, Olivier Fliche, Su Yang (2020) *Governance of Artificial Intelligence in Finance* [Online], Fintech-Innovation Hub, ACPR. Verfügbar unter https://acpr.banque-france.fr/sites/default/files/medias/documents/20200612_ai_governance_finance.pdf (Abgerufen am 30 Juli 2022).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989) „Backpropagation Applied to Handwritten Zip Code Recognition“, *Neural Computation*, Vol. 1, No. 4, S. 541–551.

Lekkala, S., Motwani, T., Parmar, M. & Phadke, A. (2021) *Emerging AI Security Threats for Autonomous Cars -- Case Studies*, arXiv [Online]. DOI: 10.48550/arXiv.2109.04865 (Abgerufen am 18 August 2022).

Leyendecker, B. & Pötters, P. (2021) *WERKZEUGE FR DAS PROJEKT- UND PROZESSMANAGEMENT: Klassische und moderne* [Online], [S.l.], Gabler. Verfügbar unter https://link.springer.com/content/pdf/10.1007/978-3-658-34724-6.pdf (Abgerufen am 17 Juni 2022).

Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C. & Tung, K.-Y. (2013) „Intrusion detection system: A comprehensive review“, *Journal of Network and Computer Applications*, Vol. 36, No. 1, S. 16–24 [Online]. DOI: 10.1016/j.jnca.2012.09.004 (Abgerufen am 18 August 2022).

Lossos, C., Geschwill, S. & Morelli, F. (2021) „Offenheit durch XAI bei ML-unterstützten Entscheidungen: Ein Baustein zur Optimierung von Entscheidungen im Unternehmen?“, *HMD Praxis der Wirtschaftsinformatik*, Vol. 58, No. 2, S. 303–320 [Online]. DOI: 10.1365/s40702-021-00707-1 (Abgerufen am 18 August 2022).

Mahdavifar, S. & Ghorbani, A. A. (2020) „DeNNeS: deep embedded neural network expert system for detecting cyber attacks“, *Neural Computing and Applications*, Vol. 32, No. 18, S. 14753–14780 [Online]. DOI: 10.1007/s00521-020-04830-w (Abgerufen am 18 August 2022).

Maik Morgenstern, Olaf Pursche & Eric Clausing (2021) „Datenschutz und Datensicherheit“, S. 102–106 [Online]. Verfügbar unter https://link.springer.com/content/pdf/10.1007/s11623-021-1398-1.pdf (Abgerufen am 5 Juni 2022).

Manoj Parmar & Amit Phadke (2022) *Bosch Aishield To Protect AI Systems and Bolster Digital Trust* [Online]. Verfügbar unter https://www.cyberdefensemagazine.com/bosch-aishield/ (Abgerufen am 17 Juli 2022).

Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B. Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss & Stefan Wrobel (2021) „Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz“ [Online]. Verfügbar unter https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (Abgerufen am 22 Juli 2022).

Mockapetris, P. V. (1983) *Domain names: Implementation specification* rfc883 [Online]. Verfügbar unter https://www.rfc-editor.org/rfc/rfc883.html (Abgerufen am 18 August 2022).

National Institute of Standards and Technology *Cybersecurity Framework | NIST* [Online]. Verfügbar unter https://www.nist.gov/cyberframework (Abgerufen am 26 Juli 2022).

Orekondy, T., Schiele, B. & Fritz, M. (op. 2019) „Knockoff Nets: Stealing Functionality of Black-Box Models“, *CVPR 2019: Proceedings.* Long Beach, CA, USA, 15.06.2019 - 20.06.2019. Los Alamitos, Washington, Tokyo, IEEE Computer Society, S. 4949–4958.

Ostwald, T. (2017) *Threat Modeling Data Analysis in Socio-technical Systems* [Online]. Verfügbar unter https://arxiv.org/pdf/1712.10243 (Abgerufen am 18 August 2022).

Peng, T., Harris, I. & Sawa, Y. (2018) „Detecting Phishing Attacks Using Natural Language Processing and Machine Learning“, *2018 IEEE 12th International Conference on Semantic Computing (ICSC 2018): Laguna Hills, California, USA, 31 January - 2 February 2018.* Laguna Hills, CA, USA, 1/31/2018 - 2/2/2018. Piscataway, NJ, IEEE, S. 300–301.

Pimentel, M. A., Clifton, D. A., Clifton, L. & Tarassenko, L. (2014) „A review of novelty detection“, *Signal Processing*, Vol. 99, S. 215–249 [Online]. DOI: 10.1016/j.sigpro.2013.12.026 (Abgerufen am 18 August 2022).

Qiu, S., Liu, Q., Zhou, S. & Wu, C. (2019) „Review of Artificial Intelligence Adversarial Attack and Defense Technologies“, *Applied Sciences*, Vol. 9, No. 5, S. 909 [Online]. DOI: 10.3390/app9050909 (Abgerufen am 18 August 2022).

Raab, R., Treu, D., Straßburg, S. & Beckmann, H. (2021) *State-of-the-Art der IT-Governance-Forschung* [Online]. DOI: 10.18420/INFORMATIK2021-122 (Abgerufen am 18 August 2022).

Renda, A. (2019) *Artificial intelligence: Ethics, governance and policy challenges : report of a CEPS task force* [Online], Brussels, Centre for European Policy Studies. Verfügbar unter https://www.ceps.eu/wp-content/uploads/2019/02/AI_TFR.pdf (Abgerufen am 18 August 2022).

Richards, S. A. (2005) „TESTING ECOLOGICAL THEORY USING THE INFORMATION-THEORETIC APPROACH: EXAMPLES AND CAUTIONARY RESULTS“, *Ecology*, Vol. 86, No. 10, S. 2805–2814 [Online]. DOI: 10.1890/05-0074 (Abgerufen am 18 August 2022).

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1985) *Learning Internal Representations by Error Propagation*, Fort Belvoir, VA, Defense Technical Information Center [Online]. DOI: 10.21236/ada164453 (Abgerufen am 18 August 2022).

Sarker, I. H., Abushark, Y. B., Alsolami, F. & Khan, A. I. (2020) „IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model“, *Symmetry*, Vol. 12, No. 5, S. 754 [Online]. DOI: 10.3390/sym12050754 (Abgerufen am 18 August 2022).

Sarker, I. H., Furhad, M. H. & Nowrozy, R. (2021) „AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions“, *SN Computer Science*, Vol. 2, No. 3, S. 1–18 [Online]. DOI: 10.1007/s42979-021-00557-0 (Abgerufen am 18 August 2022).

Savita Mohurle & Manisha Patil (2017) *A brief study of wannacry threat: Ransomware attack 2017* [Online]. Verfügbar unter https://sbgsmedia.in/2018/05/10/2261f190e292ad93d6887198d7050dec.pdf (Abgerufen am 3 August 2022).

Saygin, A. P., Cicekli, I. & Akman, V. (2003) „Turing Test: 50 Years Later“, in Moor, J. H. (Hg.) *The Turing test: The elusive standard of artificial intelligence*, Dordrecht, Kluwer Academic, S. 23–78.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. & Hall, P. (2022) *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, National Institute of Standards and Technology.

SecurityMetrics (2022) *7 Ways to Recognize a Phishing Email: Email Phishing Examples* [Online]. Verfügbar unter https://www.securitymetrics.com/blog/7-ways-recognize-phishing-email (Abgerufen am 7 August 2022).

Singh, K., Aggarwal, P., Rajivan, P. & Gonzalez, C. (2019) „Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails“, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63, No. 1, S. 453–457.

Sowa, A. (2020) *IT-PRFUNG, DATENSCHUTZAUDIT UND KENNZAHLEN FR DIE SICHERHEIT: Neue anstze* [Online], [S.l.], MORGAN KAUFMANN. Verfügbar unter https://link.springer.com/content/pdf/10.1007/978-3-658-30517-8.pdf (Abgerufen am 17 Juni 2022).

Tabassi, Elham (2022) „AI Risk Management Framework: Initial Draft - March 17, 2022“ [Online]. Verfügbar unter https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf (Abgerufen am 24 April 2022).

Tekwani, K. & Parmar, M. (2022) *Critical Checkpoints for Evaluating Defence Models Against Adversarial Attack and Robustness* [Online]. Verfügbar unter https://arxiv.org/pdf/2202.09039.pdf (Abgerufen am 18 Februar 2022).

Tencent Keen Security Lab (2019) „Experimental Security Research of Tesla Autopilot“ [Online]. Verfügbar unter https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf (Abgerufen am 1 Juli 2022).

TURING, A. M. (1950) „I.—COMPUTING MACHINERY AND INTELLIGENCE“, *Mind*, LIX, No. 236, S. 433–460 [Online]. DOI: 10.1093/mind/LIX.236.433 (Abgerufen am 18 August 2022).

Ulrike Fischer (2013) *Zeitreihenprognose in relationalen Datenbanksystemen* [Online]. Verfügbar unter https://dl.gi.de/bitstream/handle/20.500.12116/33819/21.pdf?sequence=1&isallowed=y (Abgerufen am 31 Juli 2022).

van Houdt, G., Mosquera, C. & Nápoles, G. (2020) „A review on the long short-term memory model“, *Artificial Intelligence Review*, Vol. 53, No. 8, S. 5929–5955 [Online]. DOI: 10.1007/s10462-020-09838-1 (Abgerufen am 18 August 2022).

Wang, X. & Yin, M. (2021) „Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making“, *26th International Conference on Intelligent User Interfaces*. College Station TX USA, 14 04 2021 17 04 2021. New York,NY,United States, Association for Computing Machinery, S. 318–328.

Westerlund, M. (2019) „The Emergence of Deepfake Technology: A Review“, *Technology Innovation Management Review*, Vol. 9, No. 11, S. 39–52.

Wirtz, B. W. & Weyerer, J. C. (2019) „Künstliche Intelligenz: Erscheinungsformen, Nutzungspotenziale und Anwendungsbereiche“, *WiSt - Wirtschaftswissenschaftliches Studium*, Vol. 48, No. 10, S. 4–10 [Online]. DOI: 10.15358/0340-1650-2019-10-4 (Abgerufen am 3 Juni 2022).

Wirtz, B. W., Weyerer, J. C. & Kehl, I. (2022) „Governance of artificial intelligence: A risk and guideline-based integrative framework“, *Government Information Quarterly*, S. 101685.

Wu, J. (2021) *Literature review on vulnerability detection using NLP technology* [Online]. Verfügbar unter https://arxiv.org/pdf/2104.11230 (Abgerufen am 18 August 2022).

Zhang, J. & Li, C. (2020) „Adversarial Examples: Opportunities and Challenges“, *IEEE transactions on neural networks and learning systems*, Vol. 31, No. 7, S. 2578–2593.