PLATTFORM INDUSTRIE 4.0

ERGEBNISPAPIER

Künstliche Intelligenz (KI) in Sicherheitsaspekten der Industrie 4.0

Impressum

Herausgeber
Bundesministerium für Wirtschaft
und Energie (BMWi)
Öffentlichkeitsarbeit
11019 Berlin
www.bmwi.de

Redaktionelle Verantwortung
Plattform Industrie 4.0
Bertolt-Brecht-Platz 3
10117 Berlin

Gestaltung
PRpetuum GmbH, 80801 München

Stand
Februar 2019

Bildnachweis
Gorodenkoff – Fotolia (Titel),
ipopba – iStockphoto (S. 5, S. 6),
matejmo– iStockphoto (S. 13, S. 17),
monsitj – iStockphoto (S. 23)

Diese und weitere Broschüren erhalten Sie bei:
Bundesministerium für Wirtschaft und Energie
Referat Öffentlichkeitsarbeit
E-Mail: publikationen@bundesregierung.de
www.bmwi.de

Zentraler Bestellservice:
Telefon: 030 182722721
Bestellfax: 030 18102722721

Inhalt

Präambel

Dieses Papier bezieht sich auf den Bereich IT-Sicherheit in Industrie 4.0, nicht auf IT-Sicherheit allgemein. Es spiegelt den aktuellen Stand von Anwendungen Künstlicher Intelligenz wider. Es werden keine inzwischen überholten und daher nicht mehr relevanten KI-Konzepte besprochen und ebenso wenig über futuristische Szenarien spekuliert, die heute in der KI noch nicht existieren. Es soll keine Science Fiction beschrieben, sondern auf das eingegangen werden, was heute existiert, und damit auch dessen Entwicklungsdynamik berücksichtigen. Dabei wird das Papier technisch tief genug gehen, um zumindest die prinzipiellen Wirkmechanismen von modernen KI-Konzepten und deren Einfluss auf Sicherheitskonzepte zu erklären. Wir werden durchaus beschreiben, wie ein Generative Adversarial Networks (GAN) funktioniert und welche Rolle dabei im Sicherheitsbereich dem Intrusion Detetection/Prevention System zukommt, das mit dieser Technologie gezielt getäuscht wird.

Management Summary

Dieses Papier beschreibt den aktuellen Sachstand zum Thema der Künstlichen Intelligenz (KI) in Sicherheitsaspekten der Industrie 4.0 im Frühjahr 2019. Es richtet sich an Fachleute und Entscheidungsträger aus Industrie und Politik, die ein Grundverständnis über Technologien und Anwendungen der KI für die Sicherheit vernetzter Systeme entwickeln möchten. Es verdeutlicht dabei, dass die sehr dynamische Entwicklung der KI in diesem Segment vielfältigen wirtschaftlichen Nutzen erzeugen kann, aber gleichzeitig auch als Quelle neuer Sicherheitsrisiken gesehen werden muss und somit neue Verteidigungsstrategien erfordert.

Wirtschaftliche Relevanz hat KI heute fast ausschließlich im Bereich des „Machine Learning (ML)". Deshalb ist das vorliegende Papier auf diesen Bereich der Künstlichen Intelligenz fokussiert. Alle hier diskutierten Techniken sind Teile des Machine Learning. Um ein klares Verständnis über den hier dargelegten Gegenstand sicherzustellen, wird zuerst ein kurzer historischer Überblick über KI seit der Dartmouth-Konferenz gegeben, auf der dieser Begriff geprägt wurde. Anschließend werden die heute wichtigsten Techniken des ML in seinen drei Kategorien „Supervised", „Unsupervised" und „Reinforcement" Learning skizziert. Der Schwerpunkt liegt dabei auf der Darstellung der Grundkonzepte moderner neuronaler Netztechnologien, die das Gebiet des ML heute prägen.

Anschließend werden die zentralen Sicherheitsanforderungen der Industrie 4.0 kurz rekapituliert. Auch dies erfolgt mit dem Ziel der thematischen Abgrenzung und der Vermeidung von Missverständnissen, die oft in Diskussionen über industrielle IT-Sicherheit durch Verwechselungen zwischen Security und Safety, Industrie und anderen Vertikalen, sowie Business- und Consumeraspekten entstehen.

Im zweiten Kapitel wird die Unterstützung von Sicherheitskonzepten durch KI beschrieben. Identifizierungs- und Authentifizierungsverfahren mit KI-Unterstützung werden erläutert. Techniken, die heute in anderen Bereichen wie Grenz- und Zugangskontrollen alltäglich sind, haben in vielen industriellen Anwendungsbereichen noch wenig Eingang gefunden. Ein weiteres wichtiges Thema ist die Erkennung von Anomalien in Datenströmen, gewissermaßen eine Gesundheitsüberwachung der Netzwerkaktivität im Unternehmen oder im Unternehmensverbund.

KI kann darüber hinaus Beiträge zur Erkennung von Schadsoftware leisten, wenn diese noch nicht beginnen konnte, ihre Wirkung zu entfalten. Unterschiedliche Formen der Mustererkennung können helfen, frühzeitig Risiken zu erkennen und abzuwehren. Natürlich ist der Einsatz derartiger Technologien eine hochspezialisierte Aufgabenstellung, die normalerweise in Kooperation mit einschlägig erfahrenen Partnern aus dem Software- und Dienstleistungssektor zu meistern ist. Dieses Papier ist als Hinweis zu den aktuell vorhandenen Techniken im Markt gedacht.

Im dritten Kapitel wird ein Eindruck über die dunkle Seite der KI in Sicherheitsaspekten vermittelt. Cyberangriffe auf Firmen verfolgen unterschiedliche Ziele und werden häufig mit den drei Begriffen Wirtschaftsspionage, Wirtschaftssabotage und Datendiebstahl beschrieben. Leider hat besonders die Nutzung von KI zum Angriff in der jüngsten Vergangenheit, etwa seit 2013, eine besonders dynamische Entwicklung erfahren. Wir unterscheiden in diesem Papier zwischen Cyberangriffen auf die Büro-IT, auf die Produktions-OT und die speziellen Möglichkeiten für Angriffe gegen genutzte KI-Systeme selbst. Es wird deutlich, dass grenzenlose Automatisierung auch im Sicherheitsbereich neue Risiken birgt und dass die Entfernung von menschlicher Kompetenz sorgfältig geplant werden muss, damit vermeintliche Kostenoptimierung nicht zu gravierenden Finanzrisiken führt.

Der zweite Teil dieses Kapitels ist einem völlig neuen Sicherheitsrisiko gewidmet, das auf einer erst in 2013 entstandenen neuronalen Netztechnik im Sektor des Unsupervised Learning basiert, und im September 2018 zu einer Veröffentlichung führte, die klarstellte, dass diese Technologie das Potenzial hat, die KI-basierten Schutzmaßnahmen der Erkennung von Schadsoftware zumindest in Teilbereichen außer Gefecht zu setzen. Es handelt sich um die Technologie der „GAN (Generative Adversarial Networks)“, die eigentlich gedacht ist, um in einem spieltheoretischen Gleichgewicht zwischen zwei Netzen neue Daten zu erzeugen, die von den ursprünglichen Beispielen aus einer gegebenen Grundgesamtheit durch Mustererkennung des zweiten Netzes nicht mehr unterschieden werden können. Fasst man nun ein Intrusion Detection System (IDS/IPS) als eins dieser Netze auf, beziehungsweise stellt dieses durch ein Netz unter eigener Kontrolle nach, dann kann das zweite Netz Schadsoftware erzeugen, die dieselbe verheerende Wirkung wie andere Schadsoftware aus einer gegebenen Grundgesamtheit hat, die das IDS aber als harmlos einstuft und folglich keine Abwehrmaßnahmen einleitet. Sicherheitsverantwortliche in Unternehmen sollten diese noch völlig neue Bedrohung verfolgen.

Im letzten Kapitel werden neben Schlussbemerkungen und perspektivischen Aussagen vor allem Handlungsempfehlungen an die Akteure der Industrie, also an Hersteller, Integratoren und Betreiber sowie an politisch Verantwortliche gegeben.

Dieses Papier ist durch eine neue Unterarbeitsgruppe der Arbeitsgruppe 3 der Plattform Industrie 4.0 (Sicherheit vernetzter Systeme) zum Thema KI entstanden. Wir gehen davon aus, dass KI in der Industrie und damit auch im Bereich industrieller IT-Sicherheit ein längerfristig aktuelles Thema darstellen wird und dass diese Publikation in schneller Folge Updates brauchen wird. Gleichzeitig weisen wir darauf hin, dass die vorliegende Publikation als Teil eines größeren Projektes der Plattform Industrie 4.0 zu verstehen ist, in dem die Relevanz von KI für die Industrie 4.0 in ihrer Gesamtheit diskutiert wird. Dieses Projekt wird durch die Arbeitsgruppe „Technologie- und Anwendungsszenarien“ organisiert. Unser Beitrag fokussiert sich auf Sicherheitsaspekte und soll aus diesem Blickwinkel auch einen Teilaspekt der allgemeineren Diskussion darstellen.

1. Künstliche Intelligenz: Definition und inhaltliche Abgrenzung

1.1 Historische Einordnung – Phasen der künstlichen Intelligenz

Es existiert keine allgemein akzeptierte Definition des Begriffs „Intelligenz", es gibt allenfalls Negativabgrenzungen, also Aussagen darüber, was nicht als intelligent angesehen werden kann. Die Bezeichnung „Künstliche Intelligenz" wurde von McCarthy, einem der Teilnehmer der legendären Konferenz zu KI am Dartmouth College (New Hampshire, USA, 1956) geprägt, die heute als der Beginn des KI-Zeitalters gilt. Damals wurde unter KI vor allem eine maschinelle Argumentationsfähigkeit verstanden. Im Zuge der beginnenden Hochphase der Forschungsaktivität entstanden die Sprachen LISP und PROLOG und das Konzept des Perceptrons, das als Kern der Nachbildung biologischer Gehirne gesehen wurde. Als Anfang der 70er Jahre klar wurde, dass die Erwartungen an die kurzfristig erreichbare Leistung von KI weit überschätzt worden waren, brach die Begeisterung über das Thema und mit ihr der Großteil der nationalen Forschungsfinanzierungen abrupt ab. McCarthy schrieb, dass es nicht 10 Jahre dauern würde, bis Maschinen einfache menschliche Fähigkeiten erwerben würden, sondern bis zu 500 Jahre, weil ein technischer Durchbruch im Hinblick auf die maschinelle Verarbeitungsleistung nötig sei. Der erste „KI-Winter" begann: Geldquellen versiegten und wissenschaftliche Zeitschriften wiesen Publikationen zum Thema KI kategorisch ab.

1980 begann ein neuer KI-Frühling unter einer neuen Positionsbestimmung. Die Hoffnung war, dass es attraktiv werden würde, eine Grundform von Argumentationsfähigkeit mit einem Fundus an themenspezifischem Expertenwissen zu verbinden: „Expertensysteme", die mehr Wissen akkumulieren können sollten als jede Gruppe einzelner Menschen, wurden der zentrale Gegenstand der KI. Auch diese Blütephase endete in Enttäuschung, einem zweiten KI-Winter, der vom Ende der 80er Jahre bis in den Beginn der 90er währte. Der neue Markt der Expertensysteme konnte die an ihn gestellten Geschäftserwartungen nicht erfüllen, neue Firmen brachen zusammen. Aber dennoch hatten sich in dieser Phase diverse Konzepte außerhalb des aktuellen Fokus substantiell weiterentwickelt. Insbesondere war ein Konzept zu mehrschichtigen Netzen aus Perceptrons entstanden. Rumelhart, Hinton und Williams hatten 1986 das aus den 60er Jahren bekannte Verfahren der „Backpropagation" von Schätzfehlern zur Parameteroptimierung auf mehrschichtige neuronale Netze nach etablierten Methoden der numerischen Mathematik übertragen und so das Konzept des „Deep Learning" fundiert [1]. LeCun hatte 1989 mit dem weitgehend von ihm entwickelten neuen Typ neuronaler Netze „Convolutional Neural Networks" (CNN) für flächenorientierte (Bild-)Daten die Erkennung handgeschriebener Postleitzahlen ermöglicht und damit die Postsortierung der USA automatisiert [2]: Eine Sensation. Die Basis für einen nächsten Frühling der KI war bereitet.

Mitte der 90er Jahre begann die bis heute andauernde und wahrscheinlich auf weite Sicht fortbestehende dritte Phase der KI, die sich auf Basis gereifter Erkenntnisse über realistische Nutzenerwartungen auf ein dominierendes Thema richtet: „Machine Learning (ML)". Dabei geht es um die Lösung eines breiten Spektrums realer, wirtschaftlich relevanter Probleme durch Konstruktion von Approximationsfunktionen mit nahezu beliebig komplexen Parameterkonstellationen, deren Werte aus hinreichend großen Mengen an Beispieldaten abgeleitet werden. In krassem Gegensatz zu den Orientierungen der KI in den ersten zwei Epochen liefert ML bereits heute in speziellen Bereichen eine durch vielfältige objektive Vergleiche nachgewiesene Überlegenheit maschineller kognitiver Fähigkeiten gegenüber menschlicher Leistung.

Applikationen, die mit KI-basierten Algorithmen Muster in Daten erkennen, gehören heute, im Zeitalter des mobilen Internet, zum Alltag für Milliarden von Menschen und stellen einen der größten Faktoren des Geschäftserfolgs für viele Firmen dar. Die Fundamente heutiger KI stammen teilweise aus den erfolglosen Anstrengungen früherer Epochen, besonders aber aus der explosionsartigen Steigerung der Rechenleistung in IT-Systemen aller Kategorien – vom Smartphone bis zum Internet-Mega-Datacenter – einhergehend mit der Verfügbarkeit zu extrem gesunkenen Preisen. Der technologische Durchbruch, von dem McCarthy träumte, als Hoffnung für die nächsten 500 Jahre, ist viel eher als erwartet Realität geworden. Die Rechengeschwindigkeit hat sich millionenfach erhöht, und in der IT-Industrie herrscht nun weiterhin Wettbewerb um die rasanteste Steigerung dieser maschinellen Leistungsfähigkeit. Applikationsspezifische Prozessortechnologie wird die Leistung der Systemtechnik für KI in den nächsten fünf Jahren zu heutigen Preisen nochmals um einen dreistelligen Faktor erhöhen. Der zentrale Gegenstand der KI ist heute und in der absehbaren Zukunft die Bestimmung der Parameter komplexer Approximationen mit wahrscheinlichkeitstheoretisch fundierter Basis. Damit wird die Abschätzung der Zugehörigkeit neuer Beobachtungen zu definierten Klassen oder Wertebereichen immer genauer, zuverlässiger, schneller und billiger.

1.2 Aktuelle Techniken und Einsatzfelder des Machine Learning

Machine Learning[1] besteht heute aus drei Bereichen: **Supervised Learning, Unsupervised Learning und Reinforcement Learning**. Dabei ist Supervised Learning die mit großem Abstand wirtschaftlich wichtigste Disziplin. Algorithmen nach dem Prinzip des Supervised Learning basieren auf Approximationsfunktionen mit gezielt gewählten Nicht-Linearitäten und großen Zahlen von Parametern, die in anwendungsspezifischen Architekturen konfiguriert werden. Sie benötigen hinreichend große Mengen an Daten mit Kennzeichnung (Label) des Inhalts – also beispielsweise RGB-Rasterbilder mit alphanumerischer Bezeichnung des jeweiligen thematischen Bildinhalts bezogen auf vordefinierte Klassen (z. B. 320X320 JPG/„Amsel"). Das System verarbeitet dann im Lernvorgang die eingegebenen Daten auf der Basis der aktuell vorliegenden Parameterwerte und vergleicht das Ergebnis mit dem Label. Auf Basis der dabei festgestellten Abweichung werden danach die Parameter angepasst.

Bei ML geht es im Gegensatz zu den Konzepten und Zielsetzungen der KI in den früheren Epochen nicht vorrangig um die Wiedererkennung eingegebener Daten, sondern um die Verallgemeinerung der Lerninhalte auf neue, noch unbekannte Daten, die zu den im Training spezifizierten Klassen passen. Das System lernt beispielsweise wie ein Schäferhund aussieht, nicht nur wie die Tiere auf den verwandten Trainingsbildern aussehen. Nicht das individuelle Tier, sondern die Zugehörigkeit zur Klasse wird erkannt. Das System macht eine Vorhersage (prediction) der Form: „mit 87,3%iger Wahrscheinlichkeit zeigt dieses Bild einen Schäferhund".

Supervised Learning gliedert sich in spezielle Verfahren für unterschiedliche Datentypen: Kontinuierliche im Gegensatz zu diskreten Daten, flächige Daten (Muster, Bilder) und sequentielle Daten (Sprache, Text, Audio/Video). Die Analyse kontinuierlicher Daten mit stochastischen Methoden geht auf Arbeiten von Gauss aus dem Jahr 1801 zurück und ist unter dem Namen „lineare Regression" bekannt. Sie eignet sich nicht zur Behandlung diskreter Probleme. In solchen Aufgaben geht es um die Zuordnung von Beobachtungen zu Klassen (Klassifikation). McFadden und Heckman erhielten im Jahr 2000 den Nobelpreis für Wirtschaftswissenschaften für ihre Entwicklung der Methode der „logistischen Regression", einem mit linearer Regression verwandten Prinzip zur Klassifikation. Aus heutiger Sicht kann dieses Verfahren als Grenzfall eines neuronalen Netzes gesehen werden, der keine verborgenen Schichten (hidden layers) enthält, also nur die Input- und die Output-Schicht.

Die gängigen Varianten neuronaler Netze für Supervised Learning sind

- **Multilayer Perceptron (MLP)** für Daten mit unspezifischer Struktur
- **Convolutional Neural Networks (CNN)** für 2D/3D-Mustererkennung
- **Recursive Neural Networks (RNN)** für sequentielle Datenanalyse

Ein MLP besteht aus Schichten von Perceptrons, die in Vorwärtsrichtung, von der Eingabe- bis zur Ausgabeschicht vollständig miteinander verbunden sind. Ein Netz mit nur einem hidden Layer wird als flach bezeichnet, alle anderen Strukturen sind „tiefe neuronale Netze. Die Erhöhung der Parameterzahl durch breitere Schichten (mehr Perceptrons) oder durch weitere Schichten (mehr Tiefe) führt nur dann zu Verbesserungen des Ergebnisses, wenn sie mit speziellem Design in fortgeschrittenen Netz-Architekturen einhergeht. Es ist im Allgemeinen erstrebenswert, die Parameterzahl zu beschränken, aber es gibt inzwischen auch viele Optimierungsverfahren für den Lernvorgang in allen Netzwerktypen, mit denen frühere Probleme beherrscht werden.

CNN war die erste derart fortgeschrittene Architektur. Mit einer Variante des mathematischen Verfahrens der Faltung entstand in den 90er Jahren eine Alternative zu den vollständig verbundenen Schichten der MLP, die für flächige Mustererkennung (Bilder, Videos, Komponenten daraus) deutlich bessere Ergebnisse lieferte und Konfigurationen mit mehr Schichten erlaubte. „Deep Learning" ermöglichte bemerkenswerte Fortschritte in der Bilderkennung, Maschinen lernten zu sehen. Im ImageNet Wettbewerb (1.000.000 Bildbeispiele aus 1000 Kategorien, jährlich durchgeführt 2010-2017) betrug die Fehlerrate anfangs 28 %. Das 8-schichtige CNN AlexNet der Universität Toronto senkte diesen Wert im Jahre 2012 auf 16 %. Alle folgenden Sieger waren CNN, und es entstanden drei wichtige Varianten (VGG, Inception, ResNet), mit denen die Fehlerrate auf ca. 2 % gesenkt wurde

1 Im Rahmen des vorliegenden Dokuments wird auf die Übersetzung der Fachbegriffe aus der internationalen Diskussion verzichtet, um den direkten Bezug der hier formulierten Standpunkte zu ermöglichen.

(menschliche Rate im selben Benchmark ca. 5 %). Die Zahl der Schichten wuchs auf 152. CNN sind heute in praktisch allen digitalen Kameras, inklusive Smartphones implementiert.

RNN bilden eine spezielle Gruppe von Architekturen für die Behandlungen sequentieller Daten (Text, Audio, Sprache, Video) [3]. Zwei Verfahren sind führend: Long-Short-Term Memory **(LSTM)** und Gated Recurrent Unit **(GRU)**. Das an der TU München entwickelte LSTM (Schmidhuber, Hochreiter) entstand aus Forschungsarbeiten zur Bewältigung des Vanishing Gradient Problems, das in allen neuronalen Netzen beherrscht werden muss [4]. Dabei geht es darum, dass die Anpassung der Parameter (Gewichte) der Verbindung von einer Netzschicht zur nächsthöheren dadurch erreicht wird, dass diese schrittweise in Richtung des Gradienten des Fehlers aus der Ausgabeschicht verschoben werden. Dies ist die Richtung des steilsten Abstiegs – die Verschiebung reduziert den Fehler also auf sehr effiziente Weise. Um den Vektor der Ableitungen (des Gradienten) des Fehlers in Richtung der Gewichte zu berechnen, ist für untere Schichten die Anwendung der Kettenregel der Differentialrechnung erforderlich. In diesem Vorgang entstehen Produkte aus sehr kleinen Zahlen, die für tiefe Schichten verglichen mit der Zufalls-Initialisierung der Gewichte „verschwinden". Ihre relative Größe wird zu klein, um die Gewichte zu verändern. Es entstehen ausschließlich Rundungsfehler aber keine Anpassung. In einem einfachen MLP findet deshalb oft bereits auf der dritten Schicht kein „Lernen" mehr statt. Besonders für sequentielle Probleme, in denen beispielsweise Bezüge zwischen Worten in umfangreichen Texten erkannt werden sollen, stoßen andere Architekturen an Grenzen. Praktisch jedes Smartphone enthält heue Apps mit LSTM oder GRU.

Bereits 2014 erreichte die Rechnerlast, die in Google-Rechenzentren durch KI-bezogene Anfragen (Inferenzen) generiert wurde, einen Umfang, der eine absehbare Verdopplung der weltweit notwendigen Kapazität der Cloud-RZ absehbar erscheinen ließ. Um die dafür notwendigen Kosten zu begrenzen, wurde ein applikationsspezifischer Zusatzprozessor (TPU, Tensor Processing Unit) für Standardserver entwickelt und produziert [5]. Die Nachfolgetechnologie (TPU2), mit der auch der Lernvorgang beschleunigt wird, ist derzeit in der Einführung. Die Zusammensetzung der Inferenzanfragen, nach denen die TPU1 entwickelt wurde, richtete sich zu 61 % auf MLP (5, 20M), 29 % auf LSTM (58, 52M) und 5 % auf CNN (16, 8M) – Werte in Klammern bezeichnen die häufigsten Zahlen der Schichten und Gewichte (zu lernende Netzparameter). Supervised Learning ist somit allein durch Inferenz eine massive Last moderner Cloud-RZ.

Verglichen damit spielen die anderen Sparten des ML im Hinblick auf ihre Bedeutung für den IT-Markt eher Nebenrollen. Unsupervised Learning richtet sich auf Daten ohne Labels und dient dazu, diese nach Ähnlichkeit zu strukturieren oder neue Daten zu erzeugen, die in solche Strukturen passen. Zur Datenstrukturierung existieren diverse Algorithmen. Der meistverwendete ist der recht einfache und seit langer Zeit erfolgreich eingesetzte K-means-Algorithmus, in dem iterativ Cluster erkannt werden, deren Zahl extern vorgegeben wird. Zur Erzeugung neuer Daten wurde vor fünf Jahren unter dem Namen **GAN (Generative Adversarial Networks)** ein neuer Algorithmus entwickelt, der schnell an Bedeutung gewann und gerade zu einer ganzen Klasse neuer Methoden wird [6]. Ein GAN erzeugt ein spieltheoretisches Gleichgewicht zwischen zwei gegnerisch arbeitendem KI-Systemen. Eins ist trainiert, um „echte" von „gefälschten" Daten zu unterscheiden. Das andere wandelt Rauschen in neue Daten, die als „echt" präsentiert werden. Die Wahrscheinlichkeitsverteilung für die Konstruktion neuer Daten aus Rauschen wird dabei solange iterativ angepasst, bis die neuen „gefälschten" Daten von den „echten" Daten nicht mehr unterschieden werden können.

Reinforcement Learning (RL), bestärkendes Lernen, geht auf Methoden des Operations Research (OR) aus den 50er Jahren zurück, die auch als „dynamische Programmierung" bezeichnet werden. Das Grundprinzip dieser Technik besteht in den Bellman-Gleichungen, nach denen eine optimale Entscheidung so getroffen wird, dass sie den Ist-Zustand in einen morgigen Zustand überführt, von dem bereits bekannt ist, dass er morgen optimal sein wird. Auf diese Weise kann iterativ vom Ende ausgehend der Pfad der Entscheidungen optimiert werden, und dies auch unter dem stochastischen Szenario eines Markov-Prozesses. Es geht bei RL also nicht um Erkennung, Strukturierung oder kreative Erweiterung in Datensätzen, sondern um die Erzeugung von Strategien und Plänen zur Erreichung von Zielen und dem damit verbundenen Erwerb kumulativer oder finaler Belohnungen im Prozessablauf.

KI kommt in diesem Zusammenhang erst dann zum Einsatz, wenn die Menge der möglichen Zustände und Handlungsoptionen eines solchen Prozesses zu groß ist, um die Problemlösung mit konventionellen mathematischen Methoden in der Praxis realisierbar zu machen. Beispielsweise enthält der Zustands- und Handlungsraum des GO-Spiels mehr Elemente als der Planet Erde Atome besitzt. Ein Lösungsansatz für derartige Aufgaben entsteht dadurch, dass das eigentliche Problem mit Approximationsverfahren behandelt wird, die in der Technologie für die Behandlung

neuronaler Netze verfügbar sind. „Q-Learning“ ist eine Vorgehensweise, mit der die Anwendbarkeit der etablierten OR-Techniken der „Wert-Iteration“ und der „Politik-Iteration“ auf Probleme von großer Komplexität ausgedehnt werden kann. Darüberhinausgehend ist RL aber ein Bereich extrem fortgeschrittener Verfahren für Probleme mit speziellen Eigenschaften, der erst innerhalb RL entstanden ist.

1.3 Visionen über zukünftige KI und Abgrenzung des industriell relevanten Bereichs

Aus industrieller Sicht ist KI bislang vorrangig eine neue Technologie zur Automatisierung einfacher Routinearbeit. Diese Fähigkeit ist inzwischen in vielen Berufen erreicht und neue Applikationen weiten das Einsatzfeld aus. KI hat aber in vielen Bereichen Leistungsniveaus erreicht, die menschliche Fähigkeiten weit übertreffen. Es entstehen Einsatzfelder, in denen Maschinen nicht nur kostengünstiger arbeiten, sondern auch in neue Qualitätsstufen für intellektuelle Aufgaben vorstoßen. Tiefe Neuronale Netze erkennen mit ihren teilweise 9-stelligen Parameterzahlen komplexe Muster in hochaufgelösten Bildern nicht nur schneller als Menschen, sondern in einigen Anwendungen auch viel präziser. Auch zielgerichtetes Handeln ist in exemplarisch nachgewiesenen Fällen ein Feld, das langfristig an intelligente Maschinen übergehen wird. AlphaGO hat mit Reinforcement Learning in drei Turnieren alle GO-Weltmeister vernichtend geschlagen. Die Variante AlphaZERO, hat nach vier Stunden Trainingszeit das weltweit führende Schach-Computersystem ebenso vernichtend geschlagen. Schach gilt seit 1997 als Strategiespiel, das Menschen gegen Maschinen nicht mehr gewinnen können. Moderne KI gewinnt nun mit ML sogar gegen die führenden durch Menschen programmierten Maschinen.

Es entstehen neue Anwendungsfelder in allen Bereichen, wo „übermenschliche Fähigkeiten“ durch KI entstehen: In Industrie, Privatleben, Verwaltung, Forschung, Militär und anderen. Die Vision, die KI von Anfang an hatte, geht aber konzeptionell wesentlich weiter. Bis heute ist KI nicht in der Lage, kontextbezogene Probleme zu bearbeiten, sobald der Rahmen eines einzelnen mit LSTM beherrschbaren Datenbestands überschritten wird. Dazu wäre es notwendig, dass Maschinen zumindest in einem eingeschränkten Bereich ein eigenes Weltbild aufbauen, das einen solchen Kontext darstellt. Beispielsweise ist maschinelle Übersetzung heute im Kern eine Übertragung von Vokabular. Schon die korrekte Zuordnung von Pronomen zu einem „it“ in einem englischen Satz ist eine unlösbare Aufgabe, wenn Kontextwissen erforderlich ist. „The trophy did not fit into the suitcase – it was too large“. Um dies korrekt zuzuordnen, muss man wissen, dass kleine Dinge in große passen. Der maschinelle Übersetzer wird dieselbe Zuordnung wählen, wenn der zweite Halbsatz „it was too small“ lautet. Einer der Sätze wird falsch übersetzt. Die inhaltliche Aussage ist dieselbe.

Die fehlende Fähigkeit kann man gut bei kleinen Kindern beobachten, die nicht 5000 Katzenbilder betrachten müssen, um schließlich Katzen zu erkennen. Vielmehr erscheint der menschliche Lernalgorithmus eher als „self-supervised“ – das Gehirn wiederholt Lernschritte so lange bis das Konzept verstanden ist, wendet sich dann anderen interessanten Aspekten der Welt zu und führt den Lernvorgang fort. Es ist unklar, wann es gelingen wird, eine solche Lerntechnik zu entwickeln, aber es besteht kein Grund, daran zu zweifeln, dass diese KI-Entwicklungsstufe erreicht werden wird. An diesem Punkt wird es zwei Arten von KI geben: Statistisches Parameterlernen in heutigen Kontextgrenzen und „allgemeine“ KI (auch als „starke“ KI bezeichnet).

Allgemeine KI hätte interessante Eigenschaften im Hinblick auf kontinuierliches Weiterlernen – das selbstfahrende Auto würde denselben Unfall nicht wiederholen. Gleichzeitig wird dies die Phase sein, in der ethische Fragen der KI relevant werden. Für industrielle Anwendungen ist allgemeine KI wenig hilfreich. Hier wird eine kontrollierte Umgebung angestrebt, in der sich nicht die Frage stellt, was die Maschinen im Betrieb „self-supervised“ lernen und mit welchen neuen Erkenntnissen sie das Unternehmen morgen überraschen.

Allgemeine KI ist eher ein Aspekt für philosophische Diskussionen oder ein Stoff für Science-Fiction, und es ist ein Gegenstand für Forschung, weil das Problem als solches interessant ist. Es gibt eine öffentliche Diskussion darüber, ob es zur sogenannten „Singularität“ kommen wird, in der die Welt von KI beherrscht wird und Menschen diese Welt nicht mehr verstehen, weil ihre intellektuelle Kapazität dazu nicht mehr ausreicht. Der vorliegende Artikel beschäftigt sich nicht mit solchen Fragen.

Neben der KI-Anwendung auf tatsächliche geometrische Produktdaten und tolerierbaren Abweichungen gegen technische Sollwerte können KI-Algorithmen angewendet werden, um z. B. kleinste Materialfehler in Formblechen automatisiert zu erkennen, insbesondere wenn die Qualitätsansprüche immer weiter ansteigen. Dies gilt auch für die automatisierte Erkennung von Verarbeitungsfehlern, z. B. von Pressguss-Teilen.

Der KI-Einsatz im industriellen Kontext sollte hierbei mit Augenmaß erfolgen. Im folgenden Beispiel haben entsprechende Erfahrungen aus dem industriellen Hochautomationseinsatz gezeigt [7], dass KI-Anwendungen für Produktionsabläufe unter der Kontrolle von „menschlicher Intelligenz" nach dem „human in the loop" Prinzip erfolgen sollten, um z. B. Möglichkeiten zum Bedienereingriff jederzeit sicherzustellen. Menschliches Wissen und menschliche Erfahrung werden weiterhin als erfolgskritisch eingestuft.

In diesem Sinne wurden Roboterlinien zum Schweißen bestimmter Fahrzeugbodenbleche entfernt und erfolgreich wieder durch menschliche Fachkräfte ersetzt, nachdem sich herausgestellt hat, dass rein maschinell gefertigte Schweißnähte nicht die geforderte millimetergenaue Qualität aufwiesen. Erfahrene Menschen sind während dieses speziellen Prozesses sehr schnell in der Lage, kleinste Abweichungen von Schweißnähten zu erkennen und diese sofort zu beheben.

Hier zeigte sich auf besondere Weise, dass die einmal eingerichteten vollautomatisierten Produktionslinien weiterhin in der derselben Entwicklungsstufe verharren. Weder Roboter noch KI verbessern vorhandene Prozesse. Weiterentwicklungen und Verbesserungen können nur Menschen erreichen und sollten daher für diese Aufgaben immer im Mittelpunkt bleiben.

Bei der zuvor genannten Produktionsumstellung fiel zudem deutlich auf, dass unternehmensweit das Fachwissen über die erfolgskritischen Produktionsabläufe stark rückläufig war. Es hatte über die Zeit für die Mitarbeiter keine hinreichende Gelegenheit mehr gegeben, das Fachwissen in die automatisierten Prozesse zu integrieren. Durch ein aufwendiges Verfahren mussten entsprechende Knowhow-Träger auf anderen Kontinenten gefunden werden, die dann als Keimzellen für den Wissenstransfer dienten und damit erst das Comeback der Fertigungsspezialisten ermöglichten.

Die industrielle Prozessautomation wird erst durch das Zusammenspiel von KI-Systemen mit menschlichem, übergreifendem Fachwissen sinnvoll vorangetrieben und verbessert.

1.4 Die zentralen Sicherheitsherausforderungen

Die Möglichkeiten böswilliger Angriffe auf industrielle Systeme nehmen schon auf Grund der steigenden Vernetzung im Rahmen von Industrie 4.0 zu: Vormals isolierte Anlagen werden durch Kommunikationsnetze über Ländergrenzen hinweg verbunden, die Zusammenarbeit entlang von Supply Chains wird zunehmend automatisiert. Es entstehen so zusätzliche Angriffsflächen, die es erlauben, in einem viel größeren Umfeld nach Schwachstellen zu suchen und in die Wertschöpfungskette und die damit verbundenen Unternehmen einzudringen und Schaden anzurichten. Außerdem steigen auch die Fähigkeiten potenzieller Angreifer kontinuierlich durch eine verbreiterte Verfügbarkeit von Angriffstechnologien und Knowhow, bis hin zu kommerziell buchbaren Angriffen durch kriminelle Anbieter im Darknet.

Der Anstieg vernetzter Kommunikation mit vielen beteiligten Geräten in unterschiedlichen Sicherheitsdomänen (innerhalb eines Wertschöpfungsnetzwerkes) erfordert zudem ein hohes Maß an Vertrauenswürdigkeit der Kooperationspartner untereinander. Die Einbeziehung von Geschäftspartnern in automatisierte Prozesse ist eine Herausforderung für die Koordinierung von Sicherheitsprozessen. So muss man z. B. Angriffen durch Social Engineering in allen beteiligten Sicherheitsdomänen gleichermaßen und übergreifend vorbeugen. Schwache Sicherheitsstrukturen bei einem der beteiligten Partner bedrohen automatisch auch die Sicherheit der anderen.

Dynamische und flexible I4.0-Architekturen können nicht vollständig beschrieben werden, da sich ihre Konfiguration den Anforderungen anpassen kann. Dies kann ein Anwendungsfall für KI sein. Beispiele hierfür sind eine KI-unterstützte, sich selbst organisierende Lagerverwaltung, oder eine dynamische Veränderung der Produktionsprozesskette entsprechend der Auslastung von Maschinen oder der Produktionskosten. Daher erfordern sie agile und lernfähige Sicherheitslösungen. Eine statische Anomalie-Erkennung, die z. B. auf Basis gelernter Produktionsprozesse arbeitet, würde bei oben genannten Beispielen sehr häufig Falschmeldungen ausgeben, da die Abläufe sich in kurzer Zeit grundlegend ändern können. In diesen Fällen müssen zusätzliche Metadaten in den Lernvorgang mit einbezogen werden. Auch müssen die zugrundeliegenden Lernprozesse selbst gegen Manipulation geschützt werden. Bei oben genannten Beispielen wären Angreifer bei Manipulation des Lernprozesses in der Lage, die dynamische Produktionsprozesskette derart zu manipulieren, dass kein Produkt mehr entstehen würde.

Die dazugehörigen Sicherheitsarchitekturen und Mechanismen müssen zudem den langen Lebenszyklen von Geräten und Maschinen in der Fertigung gerecht werden können. Da die Fähigkeiten von Angreifern mit der Technologieent-

wicklung zunehmen, müssen z. B. kryptographisch basierte Sicherheitsmechanismen bei Bedarf im Feld den geänderten Sicherheitsanforderungen angepasst werden (z. B. Patching).

Neben der physikalischen Implementierung einer I4.0-Produktionsumgebung existiert auch der dazugehörige „Digitale Zwilling[2]“ (vgl. Glossar der PI40). Dieser enthält prinzipiell sämtliche mit der Produktion und dem Produkt verbundenen Daten. Im „Digitalen Zwilling“ werden Produktionsprozesse geplant, simuliert, gesteuert und auch überwacht. Vieles davon geschieht auf von der physikalischen Produktion getrennten Plattformen (Cloud, Office-IT). Beide Welten kommunizieren kontinuierlich miteinander und benötigen das gleiche Security-Niveau, um zu verhindern, dass erfolgreiche Attacken das Gesamtsystem beschädigen. Es sind gleichermaßen Maschine-Maschine-Kommunikation und Mensch-Maschine-Kommunikation in einer Sicherheitsarchitektur zu berücksichtigen, wodurch die Komplexität und die Anforderungen an Sicherheitsmechanismen erhöht werden.

Als Beispiel für derartige potentielle Kommunikationsprobleme sollte man sich bewusstmachen, dass menschliche Sinnesorgane normalerweise nicht mit denselben Datenformaten und Genauigkeiten arbeiten wie die Sensorik und die Verarbeitungseinheiten von Maschinen. Menschen sehen normalerweise Bilder in 3x8-Bit RGB-Darstellung oder in 8-Bit Graustufen. Ein neuronales Netz lernt mit 16- oder 32-bit Gleitkomma-Arithmetik. Durch Addition von Rauschen im genaueren Format lassen sich Bilder erzeugen, die Menschen an Monitoren oder im Druck als identisch ansehen, die aber von Maschinen als krass unterschiedlich erkannt werden. Auf diese Weise können Angriffsszenarien entstehen, die in getrennten Prozessen zwischen Menschen und Maschinen schwer erkannt werden können.

Alle bekannten Bedrohungen aus der Industrie 3.x-Welt sind auch in der Industrie 4.0 relevant: Dort werden regelmäßig massive Cyberangriffe mit den Zielen Wirtschaftsspionage und -sabotage als auch Datenklau festgestellt, wie beispielsweise der BSI-Lagebericht für 2018 [8] und die Mitgliederbefragung vom VDMA Ende 2017 [9] zeigen. Neben Malware, Ransomware und DDoS wird im BSI-Lagebericht Schadsoftware aufgeführt, die spezielle Angriffe auf Prozessorsteuerungsanlagen (ICS) ermöglichen. Nach der VDMA-Studie führten ein Drittel der Cyberangriffe im Jahr 2017 zu Produktions- bzw. Betriebsausfällen. Häufig stehen finanzielle Ziele im Mittelpunkt, wie NotPetya gezeigt hat.

Weiterführende Informationen zu den oben genannten Herausforderungen befinden sich u. a. in folgenden Publikationen der Plattform Industrie 4.0:

- Sichere Kommunikation für Industrie 4.0 [10]
- Sichere unternehmensübergreifende Kommunikation [11]
- Sichere Identitäten [12]

Um den neuen Herausforderungen gerecht werden zu können, bedarf es zusätzlicher oder erweiterter Sicherheitsmaßnahmen, die auf Industrie 4.0-Komponenten, deren innerer Sicherheit und deren externem Kommunikationsverhalten fokussieren.

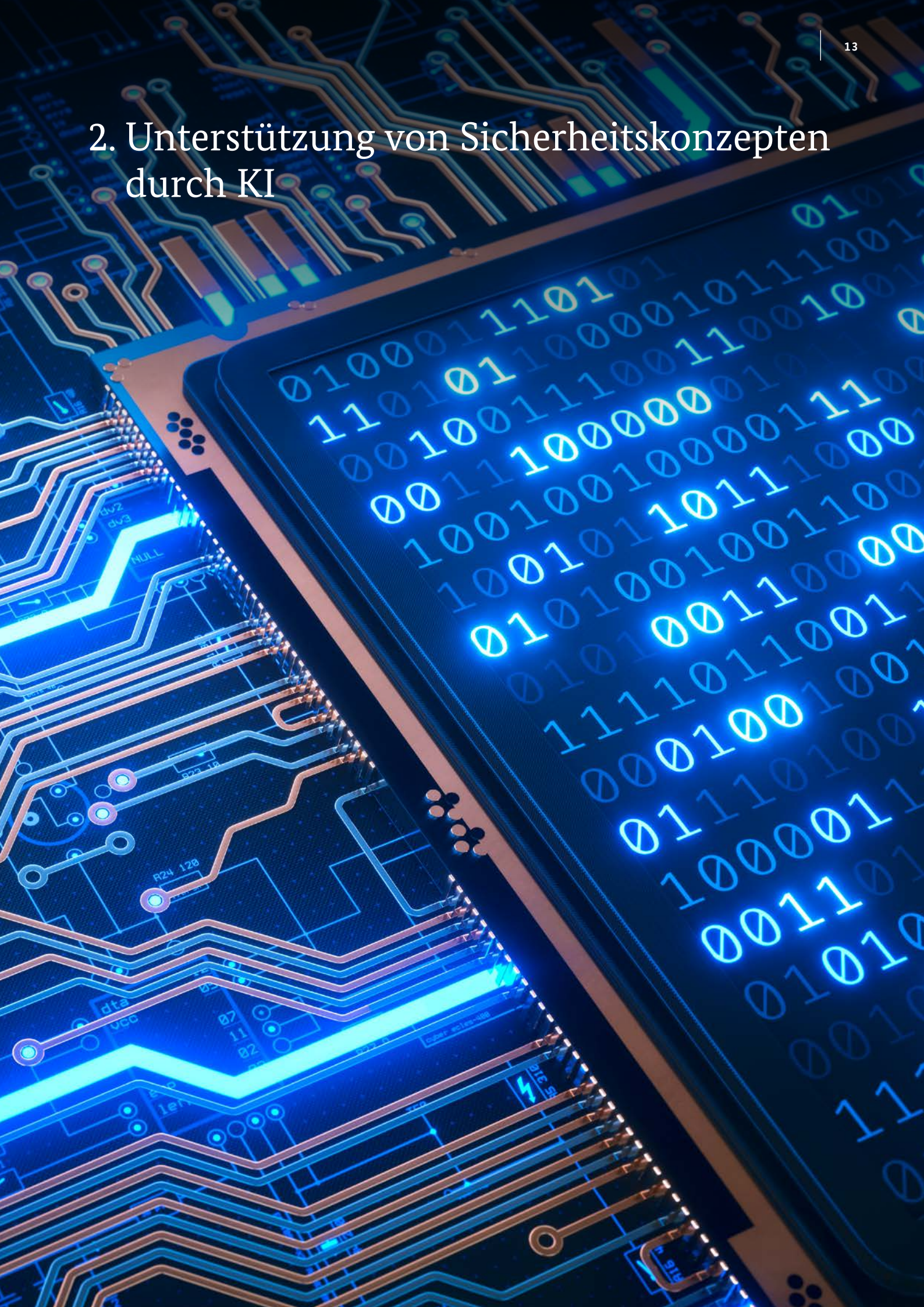
Die Fragestellungen bei der Bewältigung der Herausforderungen sind sowohl organisatorischer als auch technischer Natur.

Ein Beispiel für eine organisatorische Security-Herausforderung ist die Koordinierung übergreifender Sicherheitsmaßnahmen innerhalb eines Industrie 4.0-Wertschöpfungsnetzwerkes. Dies umfasst ggf. die Sammlung von Daten, die für Sicherheitssysteme notwendig sind, über die Unternehmensgrenzen hinaus, einschließlich von Teilen des Internets.

Bei den neuen Security-Herausforderungen technischer Natur kann KI sehr gut eingesetzt werden, z. B. um innerhalb eines Industrie 4.0-Wertschöpfungsnetzwerkes Anomalien zu erkennen und zu bewerten. Gerade weil viele Abläufe innerhalb des Wertschöpfungsnetzwerkes automatisiert (d. h. ohne den Eingriff von Menschen) und damit häufig einander ähnelnd stattfinden. D. h. die statistische Varianz nimmt ab und dadurch wird die Erkennung von abweichenden Verhalten erleichtert. Eine zusätzliche Security-Herausforderung entsteht dadurch, dass auch Angreifer von Industrie 4.0-Infrastrukturen Gebrauch von KI machen, um z. B. nach Eindringen in eine solche Infrastruktur eine geeignete Vorgehensweise bei der Analyse von vorbeilaufenden Daten zu entwickeln.

2 www.plattform-i40.de/I40/Navigation/DE/Service/Glossar/Functions/glossar.html?cms_lv2=157706

2. Unterstützung von Sicherheitskonzepten durch KI



KI ist eine Basistechnologie, die zur Automatisierung von Routineaufgaben, zur Steigerung von Geschwindigkeit und Präzision in Prozessen und zur Bewältigung von Aufgaben beitragen kann, für die vorher keine algorithmischen Lösungen existierten. Dabei können automatisierbare Routinearbeiten durchaus komplex sein und hohe fachliche Qualifikation erfordern. Mustererkennung in Medizin, Qualitätssicherung, Systemsteuerung und ähnlichen Anwendungsfeldern liefert eindrucksvolle Beispiele. Im Security-Bereich können Assistenzsysteme Fachpersonal unterstützen, einige Aufgaben komplett übernehmen, Leistungsfähigkeit von Prozessen steigern und mit maschinellem Lernen Aufgabenfelder erschließen, die für programmierte Algorithmen bislang nicht zugänglich waren. So liefert KI neue Perspektiven zur Überwindung des Fachkräftemangels und zur Verbesserung von Schutz.

2.1 Identifizierungs- und Authentisierungsverfahren mit KI-Unterstützung

Mit der Einführung des elektronischen Reisepasses am 01. November 2005 ist die Nutzung von biometrischen Daten zur Personenüberprüfung dem reisenden Bürger in Deutschland nähergebracht worden. Die Zutrittssysteme im industriellen Umfeld sind analog zu betrachten. Einige Firmen, Banken und besonders hochgesicherte Einrichtungen in Deutschland verwenden auf Biometrie-Technologie gestützt Zutrittsverfahren, z. B. um in ein Gebäude oder in einzelne Räume im Gebäude zu gelangen. Dies kann durch eine Ein-Faktor-Authentisierung erfolgen: Zentraldatenbank, vorherige Registrierung der berechtigten Person und der 1:N-Datenvergleich. Oder es kann durch eine Zwei-Faktor-Authentisierung erfolgen, mit z. B. Besitz einer Ausweiskarte und einer individuellen Eigenschaft, wie ein Fingerabdruck. Der Ausweis ist in diesem Fall üblicherweise ein Pointer auf einen Datensatz in der Datenbank und es erfolgt ein 1:1-Datenvergleich.

In Japan setzen einige Banken seit mehr als fünf Jahren in Bank-Automaten neben der PIN-Eingabe auch die Möglichkeit der Fingerabdruck-Erkennung ein, und seit diesem Jahr sogar eine Gesichtserkennung mit Audio-Begrüßung des Kunden. Nachdem die elektronische Bankkarte in den Bank-Automaten gesteckt wurde, kann der Kunde sein Geld abheben.

Eine Supermarktkette hat 2014 in Hamburg einen Pilotversuch an den Kassen durchgeführt, bei dem eine Kundenbestätigung des Einkaufs durch den Fingerabdruck verbunden mit einem automatisierten Lastschrifteinzugsverfahren erfolgt. Mit dem dahinterstehenden KI-System erscheint diese Technologie deutlich schneller, als das Bezahlen mit Bargeld oder mit einer Bankkarte oder mit dem Smart Phone.

Biometrie-Technologien zur Personenidentifikation beruhen auf Bildaufnahmen und automatisierten Bilddatenvergleichen. So wird der elektronische Bilddatensatz im Reisepass gegen die Aufnahme der Kamera im eGate verglichen. Zum Datenvergleich zwischen gespeichertem Gesichtsbild und dem Kamerabild werden KI-Systeme eingesetzt. Es könnte aber auch ein Abgleich gegen eine Suchdatei oder gegen eine „No-Fly"-Liste erfolgen, die einige Länder ständig aktualisiert herausgeben.

Neben dieser, auf individuellen Personendaten ausgerichtete KI-Anwendung werden auch verhaltensbasierte Ansätze für KI-Systeme eingesetzt, wie die Gangart des Menschen oder das Tippmuster auf der Tastatur eines PCs oder Laptops, die eine Identifizierung einer Person anstelle einer Maschine und eine Authentisierung der Person eindeutig verifiziert. Ein weiteres Beispiel ist die computergestützte Handschrifterkennung, z. B. auf Verträgen und Urkunden. Zum Trainieren des KI-Systems sind hierbei große Datensätze eine wichtige Voraussetzung für die Anwendbarkeit und die hohe Qualität der Erkennungsrate.

Bei Bild- und Videoaufnahmen können KI-basierte Algorithmen angewendet werden:

a. Zur Erkennung und Lokalisierung von Gesichtern und anderer gesuchter Merkmale in einem Bild (Treffer werden dabei durch farbige Rahmen markiert) und

b. zur Identifizierung von Personen durch Erkennung gelernter individueller Eigenschaften und Merkmale.

Die Erfassung mit Video-Überwachungskameras kann im Sekundenbereich erfolgen. Sie stellt aber im Allgemeinen keine beweisfeste Authentifizierung dar, kann aber die Zahl der durchgeführten Identifikationsprozesse dramatisch erhöhen. Verdächtige Personen können in großen Menschenmengen erkannt, in ihrem Verhalten beurteilt und zu mitgeführten Gegenständen in Beziehung gesetzt werden.

In einigen Metropolen in Europa werden beide KI-Verfahren z. B. bei der Überwachung von öffentlichen Plätzen und Einrichtungen, z. T. mit mehreren Tausend Video-Aufzeichnungsgeräten angewendet.

Einige Großstädte, wie London und Stockholm, verwenden automatisierte Fahrzeug-Kennzeichen-Erfassungssysteme, um die City-Maut verursachergerecht einzukassieren. Diese Systeme basieren ebenfalls auf KI-Systemen.

Übertragen auf Industrie 4.0 können z. B. Sensordaten, aufgenommen von Maschinen, individuelle Merkmale der Maschine und deren Betriebszustände darstellen und genutzt werden. So können beispielsweise Geräusch- und/oder Resonanzmessungen, Temperaturwerte im Betrieb, Kraftsteuerung, Wegsteuerung aber auch andere Daten von Produktionsmaschinen benutzt werden, um z. B. durch KI-Algorithmen bevorstehende Wartungs- oder Austausch-Termine zu berechnen. Insgesamt gesehen können KI-Algorithmen für drei Ziele angewendet werden:

- Vorhersagen erstellen
- Vorbeugung planen
- Maßnahmen einleiten

Auch die 2D-/3D-Bilderkennung und der KI-gestützte Datenvergleich gegen Referenzbilder können in der automatisierten Qualitätssicherung von (Zwischen-)Produkten zum Einsatz kommen, um die Qualität in der Produktion weiter zu steigern und die Überprüfungszeit und damit die Produktkosten zu reduzieren.

2.2. Anomalieerkennung in Datenströmen

Eine der bekanntesten Sicherheitsmaßnahmen zur Erhöhung der IT-Sicherheit in Firmennetzen sind Angriffserkennungssysteme, im englischen Intrusion Detection Systems (IDS) genannt, angewendet auf Computersysteme und Rechennetze. Sie können Firewalls ergänzen und in der erweiterten Ausführung als Intrusion Detection and Prevention System (IDPS) die Cyberangriffe stückweise automatisiert und aktiv verhindern.

Drei Ausprägungen werden heute am Markt angeboten: Host-basierte (HIDS), Netzwerk-basierte (NIDS) und hybride (HIDS) Angriffserkennungssysteme. Während HIDS Informationen aus Log-Dateien, Kernel-Dateien und Datenbanken analysieren, werden NIDS eingesetzt, um Daten-Pakete im Netz zu untersuchen, HIDS wiederum verbinden beide Prinzipien in einem Tool.

Für die genannten Systeme müssen zwei IDPS-Typen unterschieden werden: Die Erkennung eines Missbrauchs gegenüber der Erkennung einer Anomalie. Für die Erkennung eines Missbrauchs werden aus der Modellierung eines Angriffs spezifische Pattern extrahiert, gegen die das System systematisch durchsucht wird. Hingegen werden Anomalien identifiziert, wenn das Verhalten des Systems signifikant vom ‚Normalen' abweicht. Daher muss eine erkannte Anomalie nicht notwendigerweise einen Missbrauch darstellen.

IDPS, ergänzt mit KI-Funktionen, können gezielt mit Angriffsmustern, z. B. mit Denial of Service Angriffen (DoS) trainiert werden, umgekehrt aber auch mit sogenannten „Normalzuständen", um Anomalien zu erkennen. Im Allgemeinen stellt die sichere Erkennung dieser Normalzustände das schwierigste Problem solcher Verfahren dar, besonders dann, wenn sich normale Zustände dynamisch verändern können und das Auftreten neuer Muster normal ist.

Da klassische IDPS-Tools durch die Erzeugung von Fehlalarmen (sog. ‚False Positives') eine „Alarm-Müdigkeit" beim Sicherheitspersonal erzeugen können, ist eine Vermeidung dieser unerwünschten Erscheinung durch verhaltensorientierte Ansätze durch Einsatz von IDPS-KI-Tools möglich.

Das Anwendungsspektrum der KI reicht jedoch heute schon deutlich über diese Benutzungsvariante hinaus. Als wiederum einfaches Beispiel hierfür kann über KI auch in Kombination mit den von klassischen Suchmustern von IDPS-Tools und deren Alarmen eine Klassifikation (z. B. Typ des Angriffs, Kritikalität, echter Alarm oder Fehlalarm, ...) solcher Alarme vorgenommen werden. Mit Hilfe eines Supervised Learning-Ansatzes können die Ergebnisse im Laufe der Zeit im Betrieb der Anwendung weiter verbessert werden.

Marktprognosen erwarten, dass ab 2020 neue Technologien und Methoden, wie Analytik, maschinelles Lernen und verhaltensbasiertes Erkennen in den meisten IDPS-Tools integriert sind und am Markt angeboten werden.

Trotz dieses Fortschritts im Arsenal der Abwehrmaßnahmen sollten moderne IDPS-KI-Tools nicht als Universal-Lösung gegen Cyberangriffe betrachtet werden, schon gar nicht für alle Zukunft, sondern als neuen Meilenstein im „Hase-Igel-Wettlauf" zwischen intelligenteren Angreifern und Verteidigern.

2.3 Erkennung von Schadsoftware

Ein weiteres Anwendungsgebiet von KI liegt in der Erkennung von Schadsoftware zu einem möglichst frühen Zeitpunkt nach dem Eindringen.

Machine Learning-Techniken können verwendet werden, um Malware auf einem einzelnen Gerät sowie im Netzwerk zu erkennen. Hierzu gibt es zwei Möglichkeiten. Erstens, die Überwachung des Systems auf Anomalieerkennung entweder

a. über die Netzwerkaktivität durch einen Überwachungsserver oder einen Überwachungsdienst im Netzwerk oder

b. des einzelnen Gerätes durch die Analyse seiner Eigenschaften, wie etwa Leistungsindikatoren der Hardware und Statistiken über verschiedene Prozesse.

Zweitens kann potentielle Malware mit Klassifizierungsalgorithmen basierend auf ML analysiert werden, um eine mögliche Bösartigkeit zu identifizieren. Verschiedene Arten von Malware haben unterschiedliche „Signaturen“, die zur Kategorisierung herangezogen werden. Allerdings sind die Muster zwischen nur leicht veränderter Malware unscharf und können daher mit klassischen Methoden nicht leicht erkannt werden. ML bietet die Möglichkeit, auch diese zunächst unscharfen Unterschiede in den Mustern der Malware zu identifiziern. Die kann in einer statischen oder dynamische Analyse einer ausführbaren Datei erfolgen. In der Regel sind solche Anwendungen von KI-Maßnahmen, die in den Bereich der Hersteller von Schutzsystemen fallen und dort die Zuverlässigkeit und Aktualität von Security-Updates erhöhen.

Grundlage solcher Früherkennung ist die Tatsache, dass Schadsoftware meist aufgrund ihrer Komplexität nicht komplett neugeschrieben wird. Zumindest Code-Fragmente werden in vielen Fällen wiederverwendet und moderne ML-Techniken können Lernvorgänge ermöglichen, mit der KI verdächtigen Code erkennt, selbst dann, wenn versucht wurde, das äußere Erscheinungsbild oder das Wirkmuster zu verändern. KI kann auf diese Weise potentielle Schadsoftware erkennen, damit deren Qualität anschließend von Spezialisten überprüft werden kann. KI wirkt in diesem Fall als Mittel zur automatisierten Analyse einer großen Menge von Kandidaten und Verdachtsmomenten.

Neben der Code- und Wirkungsanalyse in kontrollierter Umgebung können weitere Merkmale gelernt werden, die sich auf die Verbreitungswege einer potentiellen Schadsoftware beziehen. Auf welchem Weg kam die betreffende Software in das Netzwerk, könnte ihr Ursprung mit anderen bekannten Risikoquellen verwandt sein? KI-gestützte Schadsoftware-Identifikation kann viele Verdachtsmomente aus gelernten Erfahrungen ableiten und so ein wertvolles Assistenzsystem bei der Konstruktion von Schutzsystemen und deren Aktualisierung darstellen. Gleichzeitig ist diese Anwendung von KI ein Beispiel für Themenbereiche, in denen menschliche Arbeit unterstützt aber nicht überflüssig wird, denn Klassifizierung von Software als schädlich ist ein sensibler Bereich, der mit erheblichen wirtschaftlichen Risiken verbunden ist. Nachvollziehbarkeit von Entscheidungen und wirtschaftliche Verantwortlichkeit erfordert heute noch menschliche Beteiligung.

3. Entstehung neuer Angriffsvektoren durch KI und Maßnahmen zu deren Abwehr

3.1 Einsatz von KI zum Angriff

Cyberangriffe auf Firmen verfolgen unterschiedliche Ziele und werden häufig mit den drei Begriffen Wirtschaftsspionage, Wirtschaftssabotage und Datendiebstahl beschrieben. Neben dem Erlangen von vertraulich eingestuften Informationen einer Firma, wie die neueste technische Entwicklung einer Maschine oder eines Produkts, stehen auch finanzielle Ziele im Mittelpunkt von Cyberangriffen, wie NotPetya gezeigt hat. Erfolgen Cyberangriffe mit KI-Unterstützung sind diese zielgenauer, präziser und wirkungsvoller zur Überwindung von Kontrollsystemen, wie auch in Kapitel 2.1 beschrieben. Die Kombination aus menschlichen und computergestützten Angriffen nutzt die Fähigkeit, aus verschiedensten Datenquellen und Kommunikationssystemen in der Büro-IT-, wie auch in der Produktions-OT-Welt Schwachstellen zu identifizieren, um wirkungsvolle Cyberangriffe zu planen und durchzuführen.

Im Folgenden werden drei Bereiche für Cyberangriffe mit KI-Unterstützung einzeln beschrieben, mit Cyberangriffen auf die Büro-IT, auf die Produktions-OT und auf genutzte KI-Systeme.

3.1.1 Cyberangriffe auf die Büro-IT

Die häufigsten Angriffe erfolgen via E-Mail und Webanwendungen. Künftig kann auch über Konferenzsysteme, wie Deep-Locker [13] zeigen konnte, Schadsoftware in die Office-IT gelangen. Die Funktionsweise von Deep-Locker wird am Ende dieses Abschnitts erläutert. Prinzipiell können zwei Grundarten von KI unterstützen Cyberangriffen unterschieden werden:

- Den mehr technisch erfolgenden Angriffen
- und den Angriffen auf organisatorische Strukturen.

Zum Teil bestehen hier auch Schnittmengen zwischen beiden Grundarten bzw. kann nicht scharf zwischen ihnen getrennt werden.

Zu den einfachsten Angriffen zählen Phishing Attacken, mit dem Versenden vom massenhaften E-Mails mit Links zu unterschiedlicher Schad-Software. Einer der bekanntesten Angriffe ist WannaCry. Zu den intelligenteren Angriffen zählen Spear-Phishing Attacken, mit der Versendung von personalisierten E-Mails mit Links z. B. zu Trojaner mit Backdoor-Funktionen und für Zero-Day-Angriffen. Bei Zero-Day-Schwachstellen handelt es sich um Software-Sicherheitsfehler, die dem Softwarehersteller bekannt sind, aber keine Patches zur Behebung der Fehler vorliegen und die von Angreifern missbraucht werden können. Bei fortgeschrittenen andauernden Bedrohungen handelt es sich meist um sogenannte Advanced Persistant Threats (APT) Attacken. Durch langanhaltende Social Engineering Methoden, die mit nachfolgenden technischen Angriffen kombiniert werden, können einzelne Personen gezielt angegriffen werden.

Aufwendige Angriffe gegen die Büro-IT verwenden zunehmend KI-Methoden. Besonders fortgeschrittene Cyberangriffe ahmen das Nutzerverhalten von Personen in Schlüsselfunktionen nach, wie die Stimmen-Imitation bei Telefongesprächen oder die genutzte Anrede am Anfang, und die Gruß-Formel am Schluss einer E-Mail, z. B. des CEOs an interne Mitarbeiter, was auch als „CEO-Fraud“ bezeichnet wird.

Deep-Locker-Angriffe sind besonders gefährlich, weil auf der Basis beliebiger vorheriger Beobachtungen des Opfers, meist öffentlich verfügbarer Informationen, individuelle geheime Schlüssel für die Verschlüsselung von Schadsoftware und geheime Auslöser für die spätere Attacke über ein entsprechend trainiertes Deep Neural Network (DNN) gebildet wird. Nach der Kompromittierungsphase des Zielsystems mit dem infiltriertem DNN und dem nahezu nicht identifizierbaren verschlüsselten Schadcode kann auf dem Zielsystem der geheime Auslöser abgewartet werden. Dies erinnert an das Prinzip des berüchtigten „Schläfers“, der lange Zeit unentdeckt auf ein bestimmtes geheimes Ereignis wartet und dann zuschlägt. Wird z. B. über eine Kamera im Zielsystem ein bestimmtes Gesicht zu einer bestimmten Zeit an einem bestimmten Ort erkannt, kann eine Attacke starten.

Es kann sich aber auch um vorbestimmte Handlungen in sozialen Netzwerken oder während einer bestimmten Webkonferenz handeln. In einen Versuch konnte gezeigt werden, wie WannaCry-Ransomware in einer Videokonferenzanwendung versteckt war, ohne dass die Malware von Antivirenprogrammen oder Sandboxing erkannt wurde.

Da sich der geheime Auslöser in nahezu jedem beliebigen Datenstrom aufhalten kann, gilt eine Entdeckung aktuell als so gut wie ausgeschlossen. Wurde die Schadsoftware aktiviert, dürfte es üblicherweise für eine Entdeckung des Angriffs zu spät sein.

3.1.2 Cyberangriffe auf die Produktions-OT

Nach einer Ende 2017 veröffentlichten Analyse vom VDMA und Basis der Befragung von Firmen in Deutschland führen ein Drittel der erfolgreichen Cyberangriffe auf Firmen in Deutschland zu Produktions- bzw. Betriebsausfällen. Häufig stehen dahinter finanzielle Ziele der Angreifer, wie Erpressung bzw. Lösegeldzahlungen.

Es gibt aber auch neue Angriffsmethoden, um an Produktionsinformationen zu gelangen. Ein Beispiel sind Reinigungs-Roboter, die angemietet wurden, um Produktionshallen bei Schichtwechsel zu reinigen. Diese vernetzten digitalen Hilfsmittel können als unerwünschte Nebenfunktion Spionage-Aufgaben verfolgen. Mittels KI-gestützter Steuerung können Spionageziele angefahren und mittels eingebauter Sensorik, beispielsweise einer Kamera, die vermeintlich zur Orientierung dient, beobachtet und analysiert werden.

3.1.3 Cyberangriffe auf das genutzte KI-System

Es besteht die Gefahr, dass ein bestehendes KI-System durch einen Cyberangriff gezielt manipuliert wird, zum Beispiel ein intelligentes IDPS in einer Firma. So können Sensor-Daten einer Maschine verändert werden, bevor diese im KI-System ankommen und verarbeitet werden. Mit manipulierten Input-Daten kann der KI-Algorithmus zu falschen Aussagen bzw. Vorhersagen gelangen. Da der Quell-Code von KI-Systemen z. T. bekannt bzw. offengelegt ist, können Angreifer den KI-Algorithmus selbst zu verändern versuchen und damit die Ergebnisse gezielt beeinflussen.

KI-basierte Systeme müssen ebenso wie klassische Systeme gegen Angriffe geschützt werden. Wichtig ist dabei zunächst ein Verständnis für mögliche Angriffsflächen von KI-Systemen. Außerdem ist seit 2013 eine neue Form von Unsupervised Learning-Technologie entstanden (GAN, beschrieben in Kapitel 3.2), mit der gegnerisch arbeitende Neuronale Netze konstruiert werden können, die trainierte Erkennungsfähigkeiten eines gegebenen anderen Netzes gezielt zerstören können.

Der folgende Abschnitt gibt einen kurzen Überblick über den momentanen Forschungsstand zu möglichen Sicherheitsschwachstellen KI-basierter Systeme, insbesondere von ML-Systemen. In der momentanen Diskussion stehen Manipulationen und Angriffe auf nicht industrielle Anwendungen im Vordergrund, wie bspw. Identifikation/Border Control, Bilderkennung für autonomes Fahren oder Spracherkennung. Die Angriffe sind jedoch prinzipiell übertragbar auf beliebige Anwendungen von ML zur Analyse im Sinne der Erkennung und Klassifikation von Mustern in einer Menge von Datenpunkten (bspw. Sensordaten). Ob und wie diese in konkreten industriellen Anwendungen ausgenutzt werden können, hängt vom jeweiligen industriellen Umfeld, Szenario und Bedrohungsmodell ab.

Eine Möglichkeit zum Angriff auf ML-Systeme zur Klassifikation großer Datenmengen (z. B. Bilderkennung) ist die Manipulation der Eingabedaten. Das Training des ML-Systems ist hier bereits abgeschlossen, d. h. das ML-System selbst ist statisch und verändert sich nicht mehr.

Die Eingabedaten können auf verschiedene Arten manipuliert werden, um Fehlklassifikationen herbeizuführen:

- Ein Angreifer kann durch das Platzieren von speziell berechneten Artefakten (Aufkleber, Graffiti) gezielt Fehler bei der Klassifikation von Verkehrsschildern herbeiführen [14]. Als Beispiel dienen Stopp-Schilder, die von den untersuchten Verfahren fehlerhaft als Richtungsanzeigen interpretiert werden. Interessant ist vor allem, dass für den Menschen eindeutige Merkmale – z. B. achteckige Form des Schilds – von dem verwendeten ML-System offensichtlich nicht ausreichend gewichtet werden.

- Andere Angriffe erzeugen ähnliche Fehler, bei denen ML-Systeme zur Gesichtserkennung durch Brillen getäuscht werden [15].

- Es ist möglich gezielt synthetische Bilddaten zu erzeugen, die von ML-Systemen vermeintlich präzise klassifiziert werden [16]. Für den menschlichen Betrachter sind lediglich abstrakte Muster oder Rauschen erkennbar, das ML-System klassifiziert die Bilder als Tiere, Früchte oder technische Geräte.

- Schließlich kann ein Angreifer korrekte Eingabedaten mit Artefakten oder Rauschen überlagern. ML-Systeme gelangen dann zu einer falschen Klassifizierung. Ein Beispiel für die Manipulation mit Hilfe von Rauschen zeigt das Bild eines Pandas, das zur Fehlklassifikation als Gibbon führt [17]. Dabei ist die menschliche Wahrnehmung der Bilder als Panda vollkommen identisch.

Natürlich können auch die Eingabedaten für klassische Algorithmen manipuliert werden. Bei der Verwendung von ML-Verfahren wird häufig eine höhere Resilienz oder Robustheit gegenüber unscharfen Eingabedaten erwartet.

Wie die obigen Beispiele zeigen, gilt dies nicht für gezielt berechnete Manipulationen. Hinzu kommt, dass einige Manipulationen vom menschlichen Benutzer nur schwer erkannt werden und Fehler nicht ohne weiteres nachvollziehbar sind. Dadurch können Gegenmaßnahmen, Analyse und Forensik erschwert werden.

Eine weitere Angriffsmöglichkeit gegen ML-Verfahren ist die Manipulation der Trainingsdaten. Dadurch wird dem ML-System von vorne herein ein fehlerhaftes Verhalten aufgeprägt. Das Verhalten kann im Sinne des Angreifers gesteuert werden.

Online-Übersetzungssysteme sind verbreitete Werkzeuge. Den Ergebnissen wird vermehrt ungeprüft Glauben geschenkt. Seit 2016 arbeitet ein prominenter Übersetzungsdienst als neuronales maschinelles Übersetzungssystem [18] und unterstützt eine große Anzahl unterschiedlicher Sprachen.

Über Verletzbarkeiten bzw. Seiteneffekte dieses Systems ist wenig bekannt. Werden Sprachen verwendet, für die das interne neuronale Netz aus wenigem und besonderem Trainingsmaterial aufgebaut wurde, können unter bestimmten Bedingungen falsche Ergebnisse entstehen, deren Sinn sich nicht erschließt. Die Ergebnisse ähneln den bizarren Mustern, die der Google Deep Dream Generator in Bildern erkennt und hervorhebt [19].

Da die Trainingsdaten und KI-Modelle in sog. „Black Boxes" implementiert wurden, ist eine Beurteilung der Qualität dieses Materials praktisch nicht möglich. Zugleich gilt die Maßgabe bei Google, soviel Trainingsmaterial wie möglich zu verwenden. Die Modelle sind darauf ausgerichtet, unter allen Umständen Ergebnisse zu produzieren, die menschlicher Sprache ähneln.

Wird diesem System ungewöhnlicher Input zur Übersetzung angeboten, wird das Ergebnis als flüssiger Text erscheinen, aber in keinem Zusammenhang mit dem Input stehen. Es kann nicht ausgeschlossen werden, dass beim Training auf Material zurückgegriffen wurde, das unzureichend qualitätsgesichert war und zudem aus Basistexten religiöser Schriften wie die Bibel, dem Koran, dem Tanach, der Tora und anderen bestand, da diese in vielen Sprachen vorliegen.

Wird Google Translate mit der sinnlosen Aneinanderreihung *„daba da ba du da bada ba du"* auf Somali konfrontiert, so wird vermutet, dass der Algorithmus auf Sätze der genannten Trainingsdaten zurückgreift. Das Ergebnis der obigen Übersetzung ins Englische lautet „At the end of the day, it is too late". Eine Kommentierung von Google ist nicht bekannt.

Zudem erlauben diese KI-Systeme, online „Verbesserungsvorschläge" durch Menschen einzubringen. Zwar sollen Gewichtungsfilter eine unsachgemäße Manipulation verhindern, aber sofern eine andere KI nach der GAN-Technologie (siehe 3.2) beginnt, einflussreich zu werden, stellt sich die Frage nach der Glaubwürdigkeit, also der Qualität dieser praktikabel nicht mehr evaluierbaren Systeme, zunehmend.

Hier hat der deutsche DeepL Übersetzungsdienst [20] eine andere Vorgehensweise gewählt, und neben eigenen, verbesserten Algorithmen zu allererst Wert auf hohe Qualität der Ausgangsdaten gelegt. Die Trainingsdaten stammen von durch professionelle Übersetzungsdienste erstellten, primär offiziellen Dokumenten aus dem Netz über einen Zeitraum von inzwischen gut zehn Jahren. Danach wurden die Übersetzungen in einer Mammutaktion weltweit von einer Community freiwilliger professioneller Helfer qualitätsgesichert. Aktuell stellen über eine Milliarde hochqualifizierter Trainingsdaten die Basis eines vertrauenswürdigen Netzes dar. Es wurde von namhaften europäischen und internationalen muttersprachlichen Journalisten als bestes Übersetzungssystem gewählt.

Ein weiterer Aspekt betrifft die Vertraulichkeit der Trainingsdaten. Ein Angreifer kann Trainingsdaten aus ML-Systemen zurückgewinnen [21]. An sich besteht das Ziel eines KI-Systems darin, z. B. ein neuronales Netz so zu trainieren, dass von den Trainingsdaten abstrahiert wird. Es sollen nicht genau die Eingabedaten wiedererkannt, sondern ein dahinterliegendes Konzept erlernt werden. Das Beispiel untersucht ein Verfahren, das ausgehend vom Training mit Textdaten (z. B. E-Mails) Vorschläge für die Komplettierung von Zeichenketten (Wörter oder Sätze) erzeugen soll. In der Trainingsdatenbank werden Geheimnisse in Form von Kreditkartennummern versteckt. Das ML-System lernt diese Geheimnisse und ein Angreifer, der das System später zur Erstellung eigener Texte nutzen kann (Black Box), ist in der Lage die Kreditkartennummern aus dem neuronalen Netz zurückzugewinnen. Obwohl die im neuronalen Netz gespeicherte Informationsmenge nicht ausreicht, um die gesamte Textdatenbank zu speichern, werden die Geheimnisse exakt wiedergegeben. Das Training des neuronalen Netzes wird abgebrochen bevor sog. Overfitting auftritt. Es zeigt sich im beschriebenen Beispiel, dass die Geheimnisse bereits in einer relativ frühen Trainingsphase (wenige Iterationen/Epochen) erlernt werden, bevor das Training üblicherweise beendet wird.

Aus den beschriebenen Beispielen ergeben sich mindestens die folgenden Punkte, die hinsichtlich der Sicherheit von ML-Systemen beachtet werden müssen:

- Die Integrität und Authentizität der Eingabedaten ist – wie auch bei klassischen Algorithmen – elementar. Die Robustheit/Abstraktion der ML-Verfahren gegenüber unscharfen Eingabedaten hilft nicht gegen gezielte Manipulation.

- Die Integrität und Authentizität der Trainingsdaten ist ebenso wichtig. Für den Anwender eines KI-Systems muss transparent sein, welche Daten vom Anbieter des KI-Systems zum Training verwendet wurden. Mindestens müssen die Kriterien zur Auswahl der Trainingsdaten offengelegt werden. Siehe hierzu auch das DeepL Beispiel oben. Die Ähnlichkeit der Vertrauensanforderungen an Trainings- und Validierungsdaten zu Anforderungen, Auswahl und Methodik bezüglich Statistikergebnissen ist nicht überraschend und sollte immer als kritischer menschlicher Filter verstanden werden.

- Für den menschlichen Anwender muss es möglich sein zu erkennen und zu überprüfen, ob sich das KI-System noch bestimmungsgemäß verhält. Die Erkenntnis bezüglich einer möglichen niedrigeren Fehlerquote von KI-Ergebnissen bezieht sich immer auf die „objektivierbare“ Erkennbarkeit von Fehlern. Diese Fähigkeit sinkt im Allgemeinen bei steigender Abstraktion der Trainingsdaten und Anwendungsfelder. Die Abgrenzung der KI-Ergebnisse zum sog. „gesunden Menschenverstand“ bezüglich Konsistenz und Wahrhaftigkeit der Ergebnisse wird jedoch zunehmend schwieriger und kann, wie später im Pkt. 3.2 (GAN) noch gezeigt wird, die Verifizierbarkeit der Ergebnisse auch ausschließen.

- Bei der Auswahl der Trainingsdaten müssen Vertraulichkeitsanforderungen berücksichtigt werden, da nicht ausgeschlossen werden kann, dass neben der angestrebten Abstraktion/Metadaten auch exakte Eingabedaten aus einem ML-System rekonstruiert werden können. Dies gilt insbesondere bei der Verarbeitung personenbezogener Daten im Hinblick die europäische Gesetzgebung zum Datenschutz (GDPR).

3.2 Einsatz von GAN-Technologie zur gezielten Überwindung von Schutzsystemen

Im Folgenden wird ein konkreter Prozess vorgestellt, mit dem ein Intrusion Detection System (IDS) oder auch ein Intrusion Protection System (IPS), durch ein gegnerisch konstruiertes Neuronales Netz außer Kraft gesetzt wird. Dabei ist die Technologie des IDS/IPS unerheblich. In der Regel handelt es sich um ein KI-gestütztes System, es könnte aber auch auf beliebigen anderen Verfahren aufbauen, Das konstruierte Netz lernt das Verhalten des IDS/IPS, bis es in der Lage ist, Schadsoftware zu erzeugen, die das IDS nicht mehr als schädlich erkennen kann. Dieser Prozess ist auch Ausdruck der vorher beschriebenen Asymmetrie zwischen Angreifer und Schutzsystem. Der Angreifer kann die aktuellste Version des Schutzsystems selbst zum Aufbau und Test seines Angriffsnetzes verwenden, während er selbst für das Schutzsystem nicht sichtbar ist. Es ist deshalb wichtig, ein IDS/IPS durch Konfiguration von Policies, die von Voreinstellungen deutlich abweichen, eine Individualisierung zu geben, die dem Angreifer die Nutzung der Asymmetrie erschwert. GAN (Generative Adversarial Networks) ist eine erst vor fünf Jahren entstandene Technologie aus dem Bereich Unsupervised Learning, die in ihrer kurzen Historie ein hohes Maß an Aufmerksamkeit gefunden hat. In der Klasse der generativen KI-Algorithmen stellt GAN die Pioniertechnologie für die direkten Verfahren mit impliziter Dichtefunktion dar. Durch GAN werden neue Beispiele einer gegebenen Grundstruktur erzeugt, ohne dass die zugrunde liegende Wahrscheinlichkeitsverteilung explizit geschätzt wird. Bemerkenswert an dieser Technik ist die Qualität der erzeugten Daten, die oft nicht mehr als gefälscht erkennbar sind. Die allgemeine Einsetzbarkeit der Methode ist allerdings derzeit noch nicht gegeben. Viele Aspekte der GAN stellen Gebiete der aktiven Forschung dar.

Die fundamentale Idee des GAN besteht darin, zwei neuronale Netze, den Generator und den Diskriminator, in gegnerischem Handeln ein stabiles spieltheoretisches Gleichgewicht (Nash Gleichgewicht) erreichen zu lassen. Der Diskriminator lernt mit Supervised Learning, „echte“ Elemente einer gegebenen Menge von neuen Elementen zu unterscheiden, die der Generator aus statistischem Rauschen erzeugt. Im Verlauf des Spiels bekommt der Generator Informationen über die Maßnahmen des Diskriminators, mit denen er seine Fähigkeit zur Unterscheidung zwischen echten und synthetischen Elementen aufrechterhält. Er nutzt diese Informationen zur Verbesserung der Parameter seiner eigenen Funktion zur Elementerzeugung. Man kann mathematisch beweisen, dass im Gleichgewicht

echt und synthetisch für den Diskriminator nicht mehr unterscheidbar sind: Er gibt im stabilen Endzustand des Spiels die Wahrscheinlichkeit, dass ein Element echt ist, sowohl für echte als auch für synthetische Beispiele mit 0,5 an. Die erzeugten Elemente tragen Merkmale als stammten sie aus der ursprünglichen Menge, obwohl sie darin nicht enthalten sind. Es gibt viele verblüffende Beispiele, in denen GANs lernen, Bilder von handgeschriebenen Ziffern, Bekleidungsstücken, Möbeln, Gesichtern, Gemälden oder Tonsequenzen bestimmter musikalischer Stilrichtungen zu erzeugen, die menschlichen Betrachtern als „echt" erscheinen.

In dem vielbeachteten Artikel „IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection" [22] wurde die neue Gefahr beschrieben, die von GAN-Technologie für Intrusion Detection Systeme (IDS) ausgeht. Dabei wird folgendes Konzept verwendet: Ein Generator verändert Schadsoftware durch Hinzufügung von Rauschen, ohne dabei deren Funktion zu verändern. Diese Veränderung wird solange weiterentwickelt, bis ein gegebenes IDS den zum Angriff geschriebenen Code nicht mehr erkennt und deshalb keine Schutzmaßnahmen einleitet. Die Schwierigkeit, die der Angreifer überwinden muss, besteht darin, dass das IDS zwar am Markt lizensiert werden kann, dass der Trainingsmechanismus des GAN aber mit diesem IDS in der Rolle als Diskriminator nicht möglich ist, weil dessen interner Programmablauf nicht beobachtet werden kann. Das IDS kann deshalb nicht als Quelle der Informationen dienen, die der Generator braucht, um das IDS zu täuschen.

Deshalb wird ein alternativer Diskriminator als neuronales Netz konstruiert, das lernt, die Handlungen des IDS zu imitieren. Das IDS liefert also die Kennzeichnungen „unbedenklich" oder „gefährlich", aus der der neue Diskriminator lernt, die Schadsoftware aus der Erzeugung des Generators als schädlich einzustufen. Es ist ein Supervised Learning-Prozess, in dem das IDS die Labels liefert, auf deren Basis die Gewichte im Diskriminator Netzwerk so eingestellt werden, dass das Verhalten des IDS imitiert wird. Diese Informationen, die im neuen Diskriminator dadurch entstehen, dass dieser lernt, sich genau so zu verhalten wie das als „Black Box" betrachtete IDS, können nun an den Generator übergeben werden.

Die Lücke im Informationsfluss für das GAN wird dadurch geschlossen. Der Aufbau des imitierten Abwehrverhaltens des neuen Diskriminators, der seine Information über Schädlichkeit vom IDS erhält, versetzt den Generator in die Lage, das für GAN charakteristische Nash Gleichgewicht im Spiel gegen das IDS zu realisieren. Er ist nun in der Lage, Schadsoftware zu erzeugen, die das IDS als ungefährlich ansieht und ohne Abwehr toleriert.

Es erscheint naheliegend, dass diese Form von konzeptionellem KI-Missbrauch, den die Autoren des IDSGAN Artikels beschreiben, nicht auf IDS beschränkt ist. Andere Formen von Sicherheitssystemen zur Erkennung abnormaler Zustände von Daten, im Systemverhalten oder bei ähnlichen Störungen, sind durch GAN basierte Angriffsszenarien nach ähnlichem Muster bedroht: Sobald der Code eines Diskriminators technisch zugänglich ist oder in geeigneter Form als Black Box so beobachtet werden kann, dass durch ein Neuronales Netz eine Imitation erzeugbar ist, steht die Informationsbasis bereit, die ausreicht, Schutzsysteme gezielt zu täuschen. Dabei spielt es keine Rolle, ob diese Systeme selbst KI-basiert oder konventionell programmiert sind.

GAN-Technologie wird in diesem Fall eingesetzt, um ein anderes Prinzip zu missbrauchen, das auch in der Diskussion für autonome Systeme schon lange präsent ist: Ein Neuronales Netz kann lernen, autonom zu handeln, wenn es ausreichend lange die Aktionen eines zu imitierenden Handelnden beobachtet und dabei dieselben externen Informationen wie dieser hat. Mit diesem Prinzip kann man auf sehr einfache Weise autonome Systeme entwickeln, die von Vorbildern Handlungen erlernen. Für Systeme wie das autonom fahrende Auto ist dies keine attraktive Perspektive, weil das Netz auch alle Fehler des Fahrers lernt, den es imitiert. Im vorliegenden Fall spielt dies keine Rolle: Aus dem als Black Box gesehenen IDS wird nur der Lernvorgang eines GAN-Diskriminators abgeleitet. Als Nebeneffekt wird die momentane Erkennungsfähigkeit des IDS erlernt. Die Qualität der so gewonnenen Fähigkeit ist dabei irrelevant, weil es bei IDSGAN ja nur darum geht, einen Weg zu finden, diese Fähigkeit durch gezielte Täuschung punktuell zu zerstören.

Es gibt bislang keine universell erfolgversprechenden Methoden, Angriffsszenarien auf der Basis von GAN-Technologie umfassend zu bekämpfen. Ratsam ist es daher, den Systemen zur automatisierten Erkennung von Anomalitäten und Eindringen von Schadsoftware nicht uneingeschränkt zu vertrauen, IDS/IPS durch Ergänzung mit Policies möglichst umfassend zu individualisieren, um Black Box-Attacken zu erschweren und möglichst mehrfache Schutzmechanismen zu implementieren.

4. Perspektive

4.1 Schlussbemerkungen

- Entwicklungen aus dem Jahr 2016 haben eindrucksvoll gezeigt, dass es mit Hilfe der jungen Disziplin „KI-Kryptographie“ möglich ist, neuartige kryptographische Verfahren einzusetzen, die keinem bekannten Krypto-Standard mehr folgen und ein hohes Maß an Dynamik besitzen. Die Methode [23] lernt, wie Kommunikation mit „gegnerischer neuronaler Kryptographie“ geschützt wird. Nähere Einzelheiten hierzu siehe Anhang.

- Erweiterung des Risikopotentials durch absehbare Fortschritte der KI im Hinblick auf Weiterentwicklung bestehender Technologie:

  Die dramatische Verbesserung der Leistung und des Preis-Leistungs-Verhältnisses der auf Numerik ausgerichteten Hardware ab 1995 war einer der Erfolgsfaktoren, die den Fortschritt der KI erst möglich machten. Doch diese war keine auf KI gerichtete Entwicklung, sondern ein Abfallprodukt, das aus den neuen Leistungsanforderungen für den boomenden Markt der 3D-Grafik-Systeme erwuchs. Die Numerik für die Koordinatentransformation zur interaktiven Handhabung von 3D-Modellen war strukturell sehr ähnlich zu den Anforderungen an Rechenleistung in der Tensorarithmentik von neuronalen Netzen. Supercomputer-Performance auf Basis von GPUs (Graphics Processing Unit) war eine Grundlage für die spektakulären Erfolge des Maschine Learning.

  Inzwischen ist ein großer Markt für applikationsspezifische Hardware im Bereich der Tensorverarbeitung entstanden, mit neuen Datenformaten und neuen Rechnerarchitekturen, die in 2019 den Markt erreichen und sowohl den Learning- als auch den Inferenzprozess massiv beschleunigen und weiter verbilligen werden. Vor allem in den großen Cloud-Mega-Rechenzentren wird diese Leistung in scharfem Wettbewerb angeboten und ist damit breit und billig in nahezu unbeschränktem Umfang verfügbar.

  Davon wird nur die Ausführung bestehender Verfahren – im positiven wie im negativen Sinn – profitieren. Es wird auch eine neue Klasse von Infrastruktur verfügbar werden, auf der sich die modernen ML-Verfahren effizient weiterentwickeln können. Trainingszeiten komplexer neuronaler Netze werden vom Wochen- in den Stundenbereich schrumpfen. Parameterzahlen werden vom dreistelligen Millionenbereich in den zweistelligen Milliardenbereich wachsen und damit die Kapazitätsgrenzen an biologischen Gehirnzellen übersteigen.

- Andere absehbare, mit KI verwandte technologiebezogene Sicherheitsrisiken: Quantum Computing und dessen aktuelle Vorstufen:

  Verschlüsselung ist eine Kerntechnologie für IT-Sicherheit in nahezu allen Anwendungsfeldern. Aufgrund der Vervielfachung der Hardwareleistung in den letzten 30 Jahren mussten Schlüssellängen vergrößert werden, denn über die Zerlegung in Primfaktoren können Codes gebrochen werden. Seit einigen Jahren entwickelt sich eine völlig neue Art von Computersystemen: Quantencomputer. Diese Systeme basieren auf den Effekten, die in der Quantenphysik beschrieben wurden, deren Existenz aber über lange Zeit bezweifelt wurde. Inzwischen ist es, zumindest in Ansätzen, möglich, mit großem Aufwand Computersysteme zu bauen, die auf solchen Effekten basieren und an denen die Existenz dieser Effekte beweisbar geworden ist. Solche Systeme können nur in der Nähe des absoluten Temperatur-Nullpunkts (einige Milli-Kelvin) und in einer Umgebung existieren, die frei von äußeren Einflüssen (beispielsweise Erschütterungen jeder Art) existieren. Sie funktionieren bisher nur für kurze Zeiträume und mit hohen Fehlerraten.

- Die Probleme, die solche Systeme lösen können, sind sehr speziell und basieren auf dem Konzept sogenannter Quantenbits, die gleichzeitig den Wert wahr und falsch annehmen können, dabei aber in Gruppen konsistent bleiben. Primfaktorzerlegung und kombinatorische Optimierungen sind Probleme, die auf solchen Systemen bewältigt werden können. Heute erreichen Quantencomputer nur Kapazitäten von wenigen Quantenbits. Nur sehr kleine Probleme können damit adressiert werden. Es besteht aber die Erwartung, dass in einigen Jahren größere Systeme realisierbar werden, die dann große Probleme mit nahezu unendlicher Geschwindigkeit lösen werden. Dies könnte zu einer existentiellen Bedrohung heutiger Verschlüsselungstechnik führen. Abgesehen vom Einsatz viel längerer Schlüssel ist noch keine Nachfolgetechnologie in Sicht, und es gibt auch keinen Plan für die Aufrüstung vorhandener Sicherheitssysteme.

4.2 Handlungsempfehlungen

Allgemein

- Strukturierte Zusammenarbeit auf dem Gebiet der Industrial Security zwischen Betreibern, Herstellern und Integratoren besitzt eine deutlich höhere Relevanz

bezüglich Resilienz gegen neuartige Angriffsvektoren, als bisher angenommen. Die Angriffstechniken schränken die Wirksamkeit klassischer IDS/IPS massiv ein. Dabei geht es einerseits um die ergänzende sicherheitstechnische Hochrüstung während der Transformationsphase von Industrie 3.x in I4.0-Umgebungen, zugleich aber auch um bessere Verteidigungsstrategien gegen diese Art maximalinvasiver Angriffe bei bereits bestehenden 3.x-Produktionsstrecken, die keine aktuellen Transformationsplanung unterliegen. Zu den bisherigen Sicherheitsüberlegungen einer sicheren Supply Chain kommen die oben beschriebenen Bedrohungen und Abwehrmöglichkeiten hinzu, die den Verantwortlichen ebenfalls zu vermitteln sind.

- Integratoren liefern mittels ihrer jeweiligen Anlagen und Maschinen neuartige KI-Unterstützung aus und vertrauen hierbei auch auf die Integrität ihrer Zulieferprodukte. Es müssen Kriterien gefunden werden, die es erlauben, die Qualität von verwendeten KI-Trainingsdaten und deren Relevanz für den jeweiligen Einsatzfall bestimmbar bzw. messbar zu machen.

- Der Ausbildung der Beteiligten kommt eine hohe Bedeutung zu. Es geht dabei um den Aufbau von Kompetenz zur Einschätzung der Verwendbarkeit von Open Source Software und Open Source Trainingsdaten. Die Zertifizierung von Open Source Material durch vertrauenswürdige Institutionen kann diesen Prozess unterstützen.

An Betreiber

- KI-Unterstützung ist grundsätzlich ein positives Leistungsmerkmal. Es sollte in der jeweiligen Implementation aber möglich sein, Aussagen zu erkennen, die mittels KI-Unterstützung entstanden sind, um diese Aussagen mit anderen Mitteln verifizierbar zu machen und so eine Täuschung der KI aufzudecken (vgl. Pandabeispiel oder Passkontrolle am Flughafen). Die Überprüfung der KI-Entscheidung könnte mit einem weiteren System erfolgen, das nachweisbar auf einer orthogonalen Metrik beruht.

- Mit geeigneten organisatorischen Maßnahmen sollten externe Einflüsse beispielsweise auf die Bilderkennung bei Qualitätssicherungsaufgaben unterbunden werden. Bei Zugangskontrollen sollte das Tragen von Brillen untersagt werden.

- KI-Unterstützung sollte einer besonderen Qualitätsprüfung unterzogen werden, in der insbesondere die zum Training verwendeten Daten kritisch auf technische Eignung und rechtliche Aspekte (IP) geprüft werden.

- KI-basierte Angriffe können
  - auf gezielt zur Täuschung von Erkennungssystemen konstruierten Mustern (Bild, Sprache, Text, ...) beruhen,
  - durch gelernte Imitation verstehen, wie Schutzsysteme temporär überwunden werden können, und
  - auf Schutzsysteme aller Arten angewendet werden – KI-basierte als auch konventionelle.

Schutzsysteme/Abwehrmaßnamen sollten daher mehrfache unabhängige Maßnahmen enthalten, um gezielte Täuschungen mit höherer Wahrscheinlichkeit zu erkennen. Mit weiteren Fortschritten der KI-basierten Angriffstechniken bleiben jedoch Restrisiken erhalten, die nur mit einschlägig spezialisiertem Personal bekämpft werden können.

An Hersteller

- Hier wird Kernwissen bei der Entwicklung und Nutzung von KI-Unterstützung verlangt. Eine intensive Ausbildung und die regelmäßige Verbindung zu R&D stellt wegen der aktuell hohen Innovationsgeschwindigkeit eine Grundforderung dar.

- Das Verständnis bezüglich der Relevanz von Trainingsdaten ist business-critical.

An Politik

- Die Politik sollte, wo nötig, dabei unterstützen, dass KMU über Chancen und Risiken der KI in der industriellen Security durch geeignete Maßnahmen aufgeklärt werden. Die Verantwortung hierfür liegt jedoch bei den KMU selbst.

- Aufgrund der sicherheitspolitischen und industriepolitischen Tragweite sollte geprüft werden, durch geeignete Fördermaßnahmen eine besonders für diese Sicherheitsanforderungen konzipierte geförderte Zusammenarbeit zwischen Betreibern, Herstellern und Integratoren zu initiieren, um bezüglich der neuartigen Bedrohungslage an geeigneten Prävention zu arbeiten, diese in gemeinsamen Feldversuchen zu erproben und praxistaugliche Ergebnisse zu liefern.

5. Referenzen

[1] **Rumelhart, D. E., Hinton, G. E., and Williams, R. J.** (1986b). *Learning representations by backpropagating errors.* Nature, 323, 533–536.

[2] **LeCun, Y., Jackel, L., Boser, B., and Denker, J.** (1989). *Handwritten digit recognition: Applications of neural network chips and automatic learning.* IEEE Communications Magazine, 27(11), 41–46.

[3] **Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua** (2014). *"Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation"*. arXiv:1406.1078

[4] **Hochreiter, S., Schmidhuber, J.** (1997), *Long short-term memory*, Neural computation 9 (8), 1735–1780.

[5] **Jouppi, Norman et al.** (2017), *In-Datacenter Performance Analysis of a Tensor Processing Unit*, ACM Digital Library.

[6] **Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio** (2014), *Generative Adversarial Networks*, arXiv:1406.2661 stat.ML.

[7] www.tagesspiegel.de/themen/reportage/kuenstliche-intelligenz-toyota-feuert-die-roboter/23821418.html

[8] www.bsi.bund.de/DE/Publikationen/Lageberichte/lageberichte_node.html

[9] www.vdma.org/documents/15012668/22538766/Grafik_PI_Industrial_Security_2017-11-29_1512390672976.pdf/b94c55dc-5b8f-44f1-ad03-1b7628499e21

[10] www.plattform-i40.de/I40/Redaktion/DE/Downloads/Publikation/sichere-kommunikation-i40.pdf?__blob=publicationFile&v=6

[11] www.plattform-i40.de/I40/Redaktion/DE/Downloads/Publikation/sichere-unternehmensuebergreifende-kommunikation.pdf?__blob=publicationFile&v=10

[12] www.plattform-i40.de/I40/Redaktion/DE/Downloads/Publikation/sichere-identitaeten.pdf?__blob=publicationFile&v=11

[13] https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/

[14] **Anh Nguyen et al.**, *"Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images"*, https://arxiv.org/abs/1412.1897

[15] **Mahmood Sharif et al.**, *„Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition"*, www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

[16] **Kevin Eykholt et al.**, *„Robust Physical-World Attacks on Deep Learning Models"*, CVPR 2018, https://arxiv.org/abs/1707.08945

[17] **Ian J. Goodfellow et al.**, *"Explaining and harnessing adversarial examples"* https://arxiv.org/pdf/1412.6572.pdf

[18] www.blog.google/products/translate/higher-quality-neural-translations-bunch-more-languages/

[19] https://vimeo.com/132462576

[20] https://deepl.com/

[21] **Nicholas Carlini et al.**, *„The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets“*, https://arxiv.org/pdf/1802.08232.pdf

[22] **Zilong Lin, Yong Shi, Zhi Xue** (2018), *IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection*, arXiv:1809.02077 cs.CR

[23] **Martın Abadi and David G. Andersen**: https://arxiv.org/pdf/1610.06918v1.pdf, 2016.

[24] http://deeplearning.stanford.edu/tutorial/supervised/OptimizationStochasticGradientDescent/

6. Anhang

6.1 Beispiel Grenzkontrolle mittels KI

Zunächst wurde das **Gesicht** als Datensatz elektronisch im Pass-Buch gespeichert, gefolgt von **zwei flachen Fingerabdrücken**, die ab 2009 im Pass aufgenommen wurden. Am Frankfurter Flughafen wurden zudem automatisierte Grenzkontrollen (sogenannte eGates) mit vorher gespeicherten **Iris**-Bilddaten in 2008 mit freiwilligen Nutzern mit dem EasyPass Registered Traveller Programm von der Bundespolizei getestet, die ab 2013 in Frankfurt gegen Gesichtserkennungssysteme ausgetauscht und an vier weiteren deutschen Großflughäfen installiert wurden. Vier Kernforderungen können damit erfüllt werden:

- Wiederkehrende monotone Arbeiten können von computergestützten Systemen in der immer gleichbleibenden Qualität ausgeführt werden,
- die Geschwindigkeit, mit der die Arbeit durch ein computergestütztes System ausgeführt wird, ist um ein Vielfaches schneller als bei der Ausführung durch eine Person,
- das computergestützte System ist, im Vergleich zu einer Person, in der Lage, kleinste Unterschiede der biometrischen Merkmale zu erfassen und
- computergestützte Systeme bieten, in Verbindung mit einer Datenbank, die Möglichkeit, die biometrischen Merkmale vor dem Zugriff von Dritten zu schützen.
- Diese Kernforderungen aus dem Bereich der Identifizierungs- und Authentisierungsverfahren auf Prozesse in der Industrie 4.0 zu transformieren, könnte bedeuten, KI-gestützt
- Monotone wiederkehrende Arbeiten mindestens gleichbleibender Qualität auszuführen,
- Erhöhung der Taktgeschwindigkeit und damit der Produktivität bzw.
- Qualitätsverbesserung durch sichere Erkennung von Produkt-Merkmalen und
- IP Schutz durch die Möglichkeit zu erreichen, kritische Produkt-Merkmale vor dem Zugriff von Dritten zu schützen

Heute sind in Deutschland mehr als 200 und in Europa mehr als 1000 eGates im Einsatz, weltweit geben mehr als 130 Staaten biometrische Pässe aus und mehr als 50 Staaten nutzen automatisierte Grenzkontrollen, überwiegend an Flughäfen, aber auch an Landübergängen. So überqueren tägliche mehr als 700.000 Reisende die Grenze zwischen China und Hong Kong sowie Macao, überwiegend über eGates. Auch die Brücke zwischen Singapur und Malaysia zählt täglich mehr als 100.000 Grenzübertritte, die meistens mit biometrischer Erkennung durchgeführt werden. Diese Systeme führen zu einer deutlichen Beschleunigung der Abläufe und ein möglicher Ermüdungseffekt des Faktor Mensch kann ausgeschlossen werden. Während der Grenzpolizist sich nur etwa 10 bis 20 Bilder von gesuchten Personen merken kann, können Computer mehrere Tausend Bilddaten speichern und verarbeiten.

In der Türkei werden seit 2013 alle Patienten z. B. in Krankenhäusern anhand der Venen-Muster in der Handfläche authentifiziert. Dies dient nicht nur zur Personen-Überprüfung, sondern auch um Vertauschungen von Patienten in großen Krankenhauskomplexen zu vermeiden.

6.2 Erläuterungen zu: “Learning to protect communcations with adversarial neural Crytography”

Im Brain Projekt wurden zwei KI-Instanzen erstellt, die nach Abschluss einer Trainingsphase ihr eigenes, sicheres Kryptosystem erschufen, um miteinander zu kommunizierten und dabei Angriffen einer dritten KI-Instanz standhielten, obwohl diese von Anfang an die Kommunikation ebenfalls „mithören“ konnte. Der Ansatz basiert u. a. auf der Optimierung nach SGD (Stochastic Gradient Descent) [24].

Bei der zugrunden liegenden Technologie lernen neuronale Netze geheime Schlüssel zu verwenden, um Informationen vor anderen neuronalen Netzen zu schützen und Vertraulichkeit in en in Bezug auf Gegner zu gewährleisten. In diesen Technologien werden den neuronalen Netzen keine spezifischen kryptographischen Algorithmen vorgeschrieben, sondern gegensätzlich trainieren End-to-End-Erkennung trainiert. Die neuronalen Netze lernen grundsätzliche Formen der Ver- und Entschlüsselung sowie die Verwendung zur Erreichung von Vertraulichkeitszielen. Nach ca. 7.000 Kommunikationsschritten haben die beiden KI-Instanzen sich soweit synchronisiert, dass das dritte KI-System keine Erfolgstreffer mehr hat, weder durch die Nutzung des vortrainierten DNN noch durch Erraten. Nach weiteren 6.000

Kommunikationsschritten ist nach Meinung der Forscher eine externe Instanz nicht mehr in der Lage, die permanent wechselnden Sicherheitskriterien zu „knacken“. Ob diese Form der KI-Kryptographie nach der aktuell gültigen Klassifizierung als starke Kryptographie einzustufen ist, kann aktuell nicht sicher beurteilt werden. Vor Einzug dieser Technologie im zertifizierbaren, kommerziellen Bereich müssen daher erst neue Prüf- und Zertifizierungsschemata entwickelt werden.

Einschätzung: Unabhängig von einem kontrollierten zw. kontrollierbaren Einsatz ist die unkontrollierte Nutzungsmöglichkeit bereits jetzt gegeben und für Sicherheitsverantwortliche eine Herausforderung. Sofern Malware diese Methode zur internen Kommunikation mit eigenen Komponenten, wie z. B. mit externen Control Centern, einsetzt, kann einer Firewall im Sinne einer Maskerade ein gültiges Protokoll gezeigt werden, in dem KI-verschlüsselten Information als syntaktischer gültiger Inhalt präsentiert wird und von IDS/IPS System nicht als Angriff erkannt bzw. entschlüsselt werden. Gleiches gilt z. B. für neue Hash-Methoden, die nur von den beiden beteiligten KI-Instanzen erzeugt und verifiziert werden können bzw. auch keinem bekannten Format entsprechen müssen. Es darf bereits jetzt angenommen werden, dass diese Form „Proprietärer Security“, ähnlich GAN, in bestimmten Segmenten hohes Entwicklungspotenzial aufweist.

AUTOREN

Markus Heintel, Siemens AG | Dr. Detlef Houdeau, Infineon Technologies AG | Dr. Wolfgang Klasen, Siemens AG | Dr. Bernd Kosch, Fujitsu Technology Solutions GmbH | Dr. Michael Schmitt, SAP SE | Thomas Walloschke, Fujitsu Technology Solutions GmbH | Dr. Thomas Wille, NXP Semiconductors Germany GmbH

Diese Publikation ist ein Ergebnis der Unterarbeitsgruppe der Arbeitsgruppe 3 der Plattform Industrie 4.0 (Sicherheit vernetzter Systeme).

www.plattform-i40.de