Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Künstliche Intelligenz und IT-Sicherheit

Bestandsaufnahme und Lösungsansätze

WHITEPAPER

J. Müller-Quade et al.
AG IT-Sicherheit,
Privacy, Recht und Ethik

Inhalt

Zusammenfassung 3

1. Einleitung 4

2. KI unterstützt IT-Sicherheit 6

3. Herausforderung: Dual Use-Potenzial von KI 14

4. Sicherheit und Schutz von KI-Systemen 17

5. Lösungsansätze 20

Autoren und Redaktion 23

Literatur 24

Zusammenfassung

Technologien, die auf Künstlicher Intelligenz (KI) basieren, durchdringen zunehmend alle Lebensbereiche. Ihr Beitrag zu einer verbesserten Sicherheit von IT-Systemen wie auch die Sicherheit von KI-Systemen selbst sind essenziell, damit Bürgerinnen und Bürger, Unternehmen, Politik und Behörden die Vorteile der fortschreitenden Digitalisierung nutzen können.

KI-Systeme spielen zukünftig eine wichtige Rolle zur Erhöhung der IT-Sicherheit. Methoden des maschinellen Lernens können beispielsweise eingesetzt werden, um die Fähigkeit von Angriffserkennungssystemen zu verbessern oder um in Netzwerken normale von verdächtigen Aktivitäten zu unterscheiden. KI-Systeme können Fachkräfte für IT-Sicherheit wirksam unterstützen und damit kurzfristig die Auswirkungen des Fachkräftemangels in der IT-Sicherheit abmildern (siehe Kapitel 2).

KI-Technologien bergen jedoch auch im Bereich der IT-Sicherheit ein Dual-Use-Potenzial (siehe Kapitel 3). Methoden des maschinellen Lernens, mit denen sich bisher unbekannte Sicherheitslücken in Netzwerken oder Softwaresystemen identifizieren lassen, können auch von Angreifern verwendet werden. Diese können mithilfe von Methoden und Verfahren der KI ihre Angriffsstrategien optimieren oder neue Bedrohungen entwickeln. Das Bedrohungsrisiko sollte zwar nicht überzeichnet werden, ist jedoch eine zusätzliche Motivation dafür, sich auf diesem Einsatzfeld der KI einen technologischen Vorsprung zu erarbeiten und als Entwickler und Anwender ein Bewusstsein für das Dual-Use-Potenzial zu entwickeln.

KI-Systeme werden zunehmend in Prozesse integriert, bei denen Sicherheit und Datenschutz eine zentrale Rolle spielen. Deshalb gilt es, die KI-Systeme selbst vor Angriffen zu schützen, ihre Robustheit gegenüber möglichen Manipulationen zu erhöhen und entsprechende Schutzmaßnahmen zu implementieren (siehe Kapitel 4).

Angesichts der neuen Dynamik, die KI-Systeme in den Bereich der IT-Sicherheit bringen, ergeben sich unterschiedliche Handlungsfelder – von der Unterstützung kleinerer und mittlerer Unternehmen über den Kompetenzaufbau zu KI und IT-Sicherheit bis hin zu Entwicklung und Design der Systeme selbst. Die Autorinnen und Autoren dieses Papiers bringen für diese Handlungsfelder erste Lösungsansätze ein (siehe Kapitel 5), die sie in einem nächsten Schritt ausdifferenzieren und weiterentwickeln werden.

1. Einleitung

Durch Wirtschaftsspionage, Sabotage oder Datendiebstahl entstand 2017 in Deutschland ein Schaden in Höhe von 55 Milliarden Euro, so eine Studie des Verbands der Informations- und Telekommunikationsbranche (Bitkom 2017: 5). Größere öffentliche Aufmerksamkeit erlangen zumeist nur Angriffe auf kritische Infrastrukturen oder staatliche Einrichtungen wie etwa der Hackerangriff „Wanna Cry" auf die Deutsche Bahn im Mai 2017 oder die Attacke auf den Bundestag, bei dem Ende 2018 sensible Daten von Abgeordneten abgegriffen wurden. Mit der zunehmenden Vernetzung von Unternehmen im Zuge der Industrie 4.0 wächst deren potenzielle Verwundbarkeit durch Cyberangriffe.[1]

Die Bedeutung der IT-Sicherheit im Zuge der fortschreitenden Digitalisierung ist angesichts dieser Zahlen offensichtlich. Die Bundesregierung hat hier beispielsweise mit dem Forschungsrahmenprogramm „Selbstbestimmt und sicher in der digitalen Welt 2015–2020" des Bundesministeriums für Bildung und Forschung (BMBF) Schwerpunkte für eine zukunftsgerichtete Entwicklung der IT-Sicherheit gesetzt. Mit den rasanten Fortschritten im Bereich der Künstlichen Intelligenz und des maschinellen Lernens ergibt sich auf dem Feld der IT-Sicherheit eine neue Dynamik. So rechnen 27 Prozent der Unternehmen mit potenziell gefährlicheren Hackerangriffen und Spionageaktivitäten durch den Einsatz von KI und Lernenden Systemen (IDG 2018: 23). Doch KI-Systeme lassen sich auch für intelligente Sicherheitslösungen nutzen, mit deren Hilfe beispielsweise die Detektion von Angriffen beschleunigt werden kann. Darüber hinaus gilt es, KI-Systeme selbst vor Angriffen zu schützen. Auch die Bundesregierung hat diese neue Dynamik erkannt. Sie stellt in ihrer im November 2018 verabschiedeten Strategie Künstliche Intelligenz fest, dass KI-Systeme ein hohes Niveau an IT-Sicherheit gewährleisten müssen und dass Manipulation, Missbrauch und Risiken für die öffentliche Sicherheit bestmöglich vermindert werden sollen (Bundesregierung 2018: 8).

In diesem Diskussionspapier werden die drei Dimensionen des Zusammenspiels von KI und IT-Sicherheit näher betrachtet. In Kapitel 2 wird gezeigt, wie KI-Systeme die IT-Sicherheit verbessern und Fachkräfte in diesem Bereich unterstützen können. Das anschließende Kapitel skizziert das Dual-Use-Potenzial von KI-Systemen, das es Angreifern ermöglicht, KI für bösartige oder kriminelle Zwecke einzusetzen. Kapitel 4 beleuchtet den Schutz von KI-Systemen: Wie können sie angegriffen werden und welche Schutz- und Abwehrmaßnahmen lassen sich implementieren? Davon ausgehend werden mögliche Lösungsansätze für Unternehmen, Politik und Behörden und die Forschung beschrieben.

1 Die Plattform Industrie 4.0 vertieft das Thema für die vernetzte Produktion in einer aktuellen Publikation (Plattform Industrie 4.0 2019).

Künstliche Intelligenz (KI)
Als Teilgebiet der Informatik versucht Künstliche Intelligenz, kognitive Fähigkeiten wie Lernen, Planen oder Problemlösen in Computersystemen zu verwirklichen. Zugleich steht der Begriff KI für Systeme, deren Verhalten gemeinhin menschliche Intelligenz voraussetzt. Da der Intelligenzbegriff nicht eindeutig festgelegt ist, verändert sich das Verständnis für KI jedoch abhängig vom Stand der Technik und es konnte sich noch keine allgemein akzeptierte Definition durchsetzen. Ziel der Forschung ist es, moderne KI-Systeme (Lernende Systeme) wie Maschinen, Roboter und Softwaresysteme zu befähigen, abstrakte Aufgaben und Probleme auch unter veränderten Bedingungen eigenständig zu bearbeiten und zu lösen, sodass der Mensch nicht jeden einzelnen Schritt programmieren muss. Sämtliche heute technisch umsetzbaren KI-Systeme ermöglichen eine Problemlösung in beschränkten Kontexten (z. B. Sprach- oder Bilderkennung) und zählen damit zur sogenannten schwachen KI.

Maschinelles Lernen
Maschinelles Lernen ist eine Schlüsseltechnologie der KI. Auf Basis einer großen Anzahl an Beispieldaten entwickeln Maschinen dabei mittels Mustererkennung Modelle, die auf neue, unbekannte Situationen angewendet werden können. Es ist grob zu unterscheiden zwischen überwachtem und unüberwachtem Lernen.

- Beim **überwachten Lernen** enthält der Algorithmus neben den Rohdaten auch das zu erwartende Ergebnis. Soll der Algorithmus beispielsweise lernen, einen Hund von einem Wolf zu unterscheiden, erhält er Beispielbilder von Hunden und Wölfen. Ziel ist es, dem Netz durch unterschiedliche Ein- und Ausgaben die Fähigkeit anzutrainieren, selbst Verbindungen herzustellen.
- Beim **unüberwachten Lernen** werden die Rohdaten ohne vorgegebenes Prognoseziel übergeben. Der Lernalgorithmus entwickelt selbstständig Klassifikatoren, nach denen er die Eingabemuster einteilt. Ziel ist es, in einem großen, unstrukturierten Datensatz interessante und relevante Muster zu erkennen oder die Daten kompakter zu repräsentieren. Ein Beispiel ist die Segmentierung von Kundendaten nach Zielgruppen, die man auf ähnliche Weise adressieren möchte.

Deep Learning
Deep Learning (tiefes Lernen) bezeichnet das maschinelle Lernen mit großen künstlichen neuronalen Netzen. Es handelt sich hierbei um Knotenschichten (sogenannte Neuronen), die durch eine Software realisiert und numerisch gewichtet miteinander verbunden sind. Diese Gewichtung kann während des Trainingsprozesses angepasst werden, sodass die Ergebnisse sich verbessern können. Je komplexer das Netz (gemessen an der Anzahl der Schichten, der Verbindungen zwischen Neuronen sowie der Neuronen pro Schicht), desto komplexere Sachverhalte können verarbeitet werden. Deep Learning hat in vielen Bereichen bereits bemerkenswerte Durchbrüche erzielt und wird etwa in der Verarbeitung natürlicher Sprache oder beim Erkennen von Objekten eingesetzt.

2. KI unterstützt IT-Sicherheit

KI-Systeme werden dazu genutzt, die Sicherheit von IT-Systemen zu verbessern und den Herausforderungen in unterschiedlichen Dimensionen der IT-Sicherheit zu begegnen.

2.1 Potenziale der KI für die IT-Sicherheit

KI-Verfahren, insbesondere aus dem Bereich des maschinellen Lernens, kommen bereits in heute verfügbaren IT-Sicherheitsprodukten zum Einsatz. Für Werkzeuge zum Monitoring und zur Sicherheitsanalyse, etwa von Datenströmen in der Produktion, wird regelmäßig mit KI-Unterstützung geworben. Mit der zunehmenden Vernetzung, etwa im Zuge der Entstehung des Internets der Dinge, steigt sowohl das Datenaufkommen aus unterschiedlichen Quellen als auch die Zahl potenzieller Angriffspunkte. Daraus resultiert nicht zuletzt eine große Anzahl von Sicherheitswarnungen. Das Bundesamt für Sicherheit in der Informationstechnik hat allein im Jahr 2018 rund 16 Millionen Warn-Mails versendet, um auf Gefahrensituationen aufmerksam zu machen (BSI 2018: 51). Mithilfe von maschinellen Lernverfahren können Sicherheitswarnungen priorisiert und zumindest teilautomatisiert analysiert und bearbeitet werden. So werden die personell oft knappen Cybersicherheits-Fachkräfte von dieser Routinetätigkeit entlastet und stehen vollumfänglich für die Analyse und Abwehr der relevanten Fälle zur Verfügung.

Erkennen von Angriffen

Maschinelle Lernverfahren kommen zum Einsatz, um in großen Datensets Muster, Trends und Anomalien zu identifizieren. Sie können deshalb genutzt werden, um die Performance von Angriffserkennungssystemen (siehe Kasten) zu verbessern und damit die Detektion von Angriffen zu beschleunigen. Diese Systeme benötigen aufgrund der hohen Bandbreiten heutiger und zukünftiger Netzwerke eine hohe Leistungsfähigkeit bei der Analyse großer Datenmengen.

Angriffserkennungssysteme
Ein Angriffserkennungssystem (Intrusion Detection System, IDS) ist eine fortschrittliche Sicherheitslösung, die insbesondere Unternehmen einsetzen, um Angriffe auf Computersysteme oder Netzwerke frühzeitig zu erkennen. Das Angriffserkennungssystem überwacht alle Netzwerkaktivitäten und wertet sie aus mit dem Ziel, ungewöhnlichen Datenverkehr zu erkennen. Dazu sammelt es relevante Daten, sortiert diese vor und wertet sie aus. Dies erfolgt zum einen durch eine Missbrauchserkennung: Hierbei werden Ereignisse mit Mustern und Zeichenketten (Signaturen) bekannter Angriffe verglichen, die aus vordefinierten Datenbanken stammen. Findet ein bekannter Angriff statt, kann dieser schnell mit dem entsprechenden Schweregrad bewertet werden. Unbekannte Angriffe werden hierbei allerdings nicht erkannt. Als zweite Methode nutzen Angriffserkennungs-

systeme daher die Anomalie-Erkennung. Sie identifiziert Angriffe durch ein vom Normalbetrieb abweichendes Systemverhalten – wenn etwa eine bestimmte Seitenzugriffszahl überschritten wird. Das System informiert daraufhin den Administrator.

- **Hostbasierte Angriffserkennungssysteme** werden auf den zu schützenden Systemen installiert. Sie analysieren beispielsweise Daten aus den Ereignisprotokolldateien (Logs) oder Registry-Datenbanken (Datenbank aus Konfigurationseinstellungen) der Systeme.
- **Netzwerkbasierte Angriffserkennungssysteme** beobachten den gesamten Netzwerkverkehr. Moderne Angriffserkennungssysteme sind hybrid und überwachen sowohl host- als auch netzwerkbasierte Software mit einem zentralen Managementsystem.

Das Angriffserkennungssystem dient meist als Ergänzung zu einer Firewall, die bestimmte Angriffe auf ein System nicht erkennen kann. Mit einem Angriffserkennungssystem lassen sich dank vielseitiger Technologien auch gezielt einzelne Anwendungen schützen.

Abwehrsystem

Abwehrsysteme (Intrusion-Prevention-Systeme, IPS) dienen als Erweiterung eines Angriffserkennungssystems. Sie leiten auch selbstständig Gegenmaßnahmen ein – indem sie beispielsweise den Datenverkehr einer bestimmten Quelle blockieren. Das Abwehrsystem ergänzt meist eine Firewall oder wird direkt in diese implementiert. Ein modernes IPS greift auf host- und netzwerkbasierte Sensoren zurück und schützt so ein Netzwerk und die daran angeschlossenen Systeme umfassend.

Bei einem herkömmlichen Angriffserkennungssystem müssen die zugrunde liegenden Datenbanken laufend manuell aktualisiert werden. Damit verbunden ist ein hoher Aufwand. Ein lernendes Angriffserkennungssystem kann vor dem Einsatz mit Daten über bekannte Angriffsmuster trainiert werden. Darüber hinaus kann es auch im Betrieb aus den Aktivitäten in Netzwerken oder auf Websites ein Modell normaler Aktivitäten und Datenverkehre lernen.

Konventionelle Erkennungssysteme	Erkennungssysteme mit lernender/ KI-Komponente
Software (SW) arbeitet mit starren Modellen	SW arbeitet mit adaptiven Modellen
SW erzeugt Entscheidungen basierend auf einem transparenten Regelsystem	SW erzeugt Entscheidungen auf Basis einer graduellen Bewertung
SW ist nicht lernfähig	SW lernt laufend hinzu
SW setzt Signaturen und Korrelationen gegen verschiedene Arten von Daten ein	SW lernt komplexe Muster aus einer großen Menge von Daten
Aktualisierung der SW erfolgt durch gesteuertes Update	SW befindet sich selbstständig im Updateprozess

Bei der automatisierten Abwehr von erkannten Cyber-Angriffen existieren aktuell noch wenige praktische Erfahrungen mit dem Einsatz von KI-basierten Abwehrsystemen. Zwar werden Standardfälle durchaus schon automatisiert behandelt. Die dazu eingesetzten Werkzeuge zur Orchestrierung, Automatisierung und Reaktion handeln dafür jedoch heute nach manuell eingetragenen Handlungsanweisungen, den sogenannten Playbooks, um etwa ein Mobiltelefon, von dem aus die maximale Anzahl von Authentisierungsversuchen misslungen ist, automatisch zu sperren und ggf. die darauf vorhandenen Daten aus der Ferne zu löschen.

Zukünftig wäre vorstellbar, dass ein KI-System anhand der Maßnahmen, die Fachkräfte in bestimmten Kontexten ergreifen, mögliche Reaktionen auf Cyber-Angriffe erlernt und diese in ähnlichen Folgefällen den Fachkräften im Sinne einer Entscheidungsunterstützung vorschlägt. Die Nachvollziehbarkeit der Entscheidungen (erklärbare KI) ist ein wichtiger Faktor für das Zusammenwirken der KI-Systeme mit den IT-Sicherheitsfachkräften. Während die Ergebnisse klassischer Lernverfahren oft sehr gut interpretierbar sind, ist dies bei komplexen Modellen, die auf der Grundlage großer Datenmengen gelernt werden (z. B. tiefe neuronale Netze), nicht immer gegeben und ein wichtiger Gegenstand der Forschung (Fraunhofer 2018: 6,30).

Sicherstellen der Identität

Neben der Detektion von Angriffen auf IT-Systeme können Verfahren der Künstlichen Intelligenz und des maschinellen Lernens auch für Authentisierungsverfahren eingesetzt werden. Deren Ziel ist es sicherzustellen, dass Personen oder auch Maschinen tatsächlich die von ihnen angegebene Identität besitzen. KI-basierte Systeme können bei der Authentisierung beispielsweise verwendet werden, um eine biometrische Komponente einer Zwei-Faktor-Authentisierung (siehe Kasten) zu verifizieren – etwa durch Gesichts- oder Spracherkennung. Andere Anwendungen kombinieren Benutzername und Passwort mit einer KI-gestützten Analyse der Umgebungsgeräusche (Karapanos 2018) oder setzen auf verhaltensbasierte Ansätze wie die Analyse nutzerspezifischer Tastatureingaben. Mithilfe von Verfahren der Anomalie-Erkennung können komplementär zu kryptografischen Authentisierungsprotokollen auffällige Verhaltensmuster erkannt und somit potenzielle Angreifer ermittelt werden. Damit können KI-Systeme einen wichtigen Beitrag zur Risikoanalyse leisten. Für die eindeutige Identifikation von Maschinen lassen sich beispielsweise aus spezifischen Charakteristika der in den Maschinen verbauten Halbleiter, die durch statistische Produktionsschwankungen bedingt sind, eindeutige Identitäten ableiten (Gógl 2018).

Die Möglichkeit, mithilfe von KI-Systemen biometrische Merkmale wie Sprache oder auch personenspezifische Tippmuster zu imitieren, stellt jedoch eine neue Herausforderung dar (siehe Kapitel 3): Die Systeme müssen in die Lage versetzt werden zu beurteilen, ob die Interaktion tatsächlich mit einem Menschen erfolgt oder mit einem KI-System, das nur behauptet, ein bestimmter Mensch zu sein.

Zwei-Faktor-Authentisierung mit KI-Verfahren

Die Zwei-Faktor-Authentisierung (2FA) bezeichnet den Identitätsnachweis eines Nutzers durch Kombination zweier unterschiedlicher und insbesondere unabhängiger Komponenten (Faktoren), die typischerweise durch Besitz und Wissen gekennzeichnet sind. Dabei kann es sich zum Beispiel um ein Passwort oder eine PIN (Wissen), zusammen mit einem persönlichen Sicherheitstoken wie Kreditkarte, Gesundheitskarte oder Personalausweis (Besitz) handeln.

Die Zwei-Faktor-Authentisierung ist nur dann erfolgreich, wenn beide festgelegten Faktoren zusammen eingesetzt werden und korrekt sind. Fehlt eine Komponente oder wird sie falsch verwendet, lässt sich die Zugriffsberechtigung nicht zweifelsfrei feststellen und der Zugriff wird verweigert. Dabei können auch biometrische Merkmale zum Einsatz kommen. So werden europäische ID-Karten, wie bereits heute der Personalausweis, nach neuesten in Arbeit befindlichen EU-Regularien einen biometrischen Merkmalssatz enthalten, der nur nach erfolgreicher Authentisierung vom Serversystem validiert werden kann.

Die Erkennung von Anomalien und Angriffen in Netzen und Systemen sowie biometrische Anwendungen sind etablierte Einsatzgebiete der KI in der IT-Sicherheit, für die bereits Produkte existieren. Weitere Gebiete, die sich ebenfalls für den Einsatz von maschinellem Lernen anbieten, sind unter anderem kryptografische Anwendungen – insbesondere die Seitenkanalanalyse von Chipkarten – und technische Evaluierungen. Allerdings ist die Entwicklung hier noch nicht weit fortgeschritten. Anstelle von Produkten und Werkzeugen existieren bislang nur wissenschaftliche Grundlagenarbeiten.

Seitenkanalanalysen spielen jedoch bei der Evaluierung von IT-Systemen mit kryptografischen Komponenten eine wichtige Rolle. Seitenkanalangriffe lassen sich in verschiedene Klassen einteilen. Laufzeitangriffe versuchen, etwaig vorhandene Laufzeitunterschiede auszunutzen, um den kryptografischen Schlüssel zu bestimmen, während Powerattacken den Stromverbrauch analysieren, insbesondere bei Chipkarten. Ziel ist es, maschinelle Lernverfahren zu entwickeln, die insbesondere dort eingesetzt werden könnten, wo herkömmliche statistische Methoden Muster in den Messdaten nicht oder weniger effizient identifizieren können. Da diese Anwendungen jedoch sehr spezialisiert sind, erscheint es eher unwahrscheinlich, dass in absehbarer Zeit ausgereifte Produkte zur Verfügung stehen.

Die große Stärke der heute eingesetzten KI-Systeme ist es, mithilfe maschineller Lernerfahren in riesigen Datenmengen und -strömen bei richtigem Training Anomalien zu erkennen, die Menschen oder von Menschen programmierten Systemen, die mit statischen und deterministischen Regeln arbeiten, verborgen bleiben. Demgegenüber steht die Schwäche, dass Künstliche Intelligenz zwar zwischen normal und anormal, nicht aber im Sinne der IT-Sicherheit zwischen gut und böse unterscheiden kann. Diese Differenzierung kann derzeit nur durch menschliche Analysten oder von ihnen aktiv eingespeiste Entscheidungsregeln vorgenommen werden.

Bei allen Lernenden Systemen besteht die Herausforderung, das Lernen zu überwachen, um zu vermeiden, dass sich die KI unbeabsichtigt von der Zielaufgabe wegentwickelt. Zwar stellt der Mensch eine letzte Entscheidungsinstanz dar, doch besteht das Risiko, dass er aufgrund der Komplexität einer Situation der KI blind vertraut.

2.2 Rahmenbedingungen für die IT-Sicherheit

Hohe Veränderungsgeschwindigkeit und Komplexität der Systeme

Protektiv wirkende IT-Sicherheitsmaßnahmen reduzieren proaktiv die Angriffsflächen von Netzwerken, Endgeräten, Servern, Daten, Anwendungen und Identitäten. Allerdings lässt sich ein vollständiger Schutz allenfalls theoretisch und nur für sehr kurze Zeitspannen erzielen, da sich die Bestandteile von IT-Systemen kontinuierlich und mit hoher Geschwindigkeit verändern.

Im praktischen Betrieb kommt hinzu, dass das Einspielen von Software-Updates zum Schließen von Sicherheitslücken aufgrund von vorgeschalteten funktionalen Tests, terminierten Wartungsfenstern und begrenzten Ressourcen oft erst deutlich zeitversetzt möglich ist. Vor allem im industriellen Bereich sind häufig noch Altsysteme im Einsatz, die nicht mehr update-fähig sind und deren Sicherheitslücken deshalb permanenten Charakter besitzen. IT-Systeme und Netzwerke müssen deshalb kontinuierlich überprüft werden: Gab es Angriffsversuche? Haben sich Angreifer bereits eingenistet? Bereiten diese Angreifer Missbrauch vor oder wird dieser bereits betrieben? Hier können KI-Systeme beispielsweise die Detektion von Angriffen unterstützen und beschleunigen, indem sie in riesigen Datenströmen typische Muster erlernen, erkennen und auf Anomalien hinweisen.

Mangel an Fachkräften

Für die technisch sehr anspruchsvolle kontinuierliche Überprüfung von IT-Systemen gibt es sowohl in Deutschland als auch weltweit nicht genügend Fachkräfte mit ausreichend hoher Qualifikation. In einer Studie des Capgemini Digital Transformation Institutes bekundeten 68 Prozent der befragten Unternehmen einen Mangel an Expertise im Bereich IT-Sicherheit – bei einem wachsenden Bedarf an entsprechenden Kompetenzen (Capgemini 2018: 4). KI-Systeme können IT-Fachkräfte in ihrer anspruchsvollen Tätigkeit unterstützen und damit kurzfristig dabei helfen, die quantitative Fachkräftelücke zu bewältigen. Geeignet hierfür sind beispielsweise Systeme, die auf Verfahren des maschinellen Lernens basieren und mit Daten über vergangene Angriffe trainiert wurden. Sie können den Analysten bei der Echtzeit-Auswertung von großen Mengen an digitalen Sensordaten helfen und die darin verborgenen schwachen Signale, die auf einen Cyber-Angriff hindeuten, entdecken. Unternehmerische Entscheidungen, durch den Einsatz kostengünstiger KI-basierter IT-Sicherheitslösungen die hohen Kosten für entsprechend qualifizierte Fachkräfte zu senken und die geringe Verfügbarkeit entsprechender Fachkräfte zu bewältigen, werden sich langfristig jedoch nicht als nachhaltig erweisen. Denn einerseits kann im Zuge der

fortschreitenden digitalen Vernetzung ein erhöhtes oder sogar steigendes IT-Sicherheitsrisiko angenommen werden. Andererseits kann trotz steigender Möglichkeiten zur Automatisierung von Schutz- und Abwehrmaßnahmen angenommen werden, dass auch zukünftig in vielen Fällen eine Kontrolle der Systeme und ihrer Entscheidungen durch Menschen erforderlich ist.

Asymmetrien in der IT-Sicherheit

Ein Grundproblem in der IT-Sicherheit besteht in der Asymmetrie zwischen den Fachkräften, die für den Schutz der Systeme und entsprechende Abwehrmaßnahmen verantwortlich sind, und den Angreifern. Offen ist, ob sich diese Asymmetrie durch den Einsatz von KI weiter verschärft.

■ Technische Asymmetrie

In vielen Einsatzszenarien digitaler Technologien haben die Angreifer einen entscheidenden Vorteil: Ihnen genügt eine einzige Schwachstelle in einer einzelnen Zeile Programmiercode. Fachkräfte müssen indessen einerseits bekannte Schwachstellen beheben und verteidigen und sich andererseits auch gegen noch unbekannte Schwächen – die meist sogar in zugelieferten Komponenten vergraben sind – oder Angriffsstrategien rundum prophylaktisch absichern (360°-Schutz).

■ Regulatorische Asymmetrie

Cyber-Schutzbeauftragte in Deutschland sind aus guten Gründen an deutsche und EU-Gesetze (z. B. Datenschutzrecht, Netzwerk-Informationssicherheitsrichtlinie) sowie an die für die Nutzung von KI zu erwartende Rechtsetzung und ethische Vorgaben gebunden. Für Cyberangreifer, die auch aus großer geografischer Entfernung agieren können, spielt dies keine Rolle. Wie bei anderen Formen der Kriminalität gilt auch hier: Der verbrecherischen Energie und Kreativität bei der Nutzung von KI für Cyberangriffe – sei es durch Einzeltäter, organisierte Kriminalität, Nachrichtendienste, terroristisch motivierte oder sonstige Angreifertypen – sind keine rechtlichen oder moralischen Grenzen gesetzt.

■ Asymmetrien des Zugangs

Sicherheitslösungen sind in der Regel frei am Markt – und damit prinzipiell auch für Angreifer verfügbar. Diese können ihre Systeme und Strategien gegen die Sicherheitslösungen testen und Schadsoftware sowie Angriffsstrategien entsprechend weiterentwickeln und optimieren. Grundsätzlich haben auch die Schutzbeauftragten Zugang zu Angriffstools, die sie nutzen können, um etwa Penetrationstests ihrer Rechner oder Netzwerke durchzuführen. Während sich die Angreifer am offenen Markt bedienen, ist der Zugang zu Schadprogrammen für die Verteidiger schwieriger, voraussetzungsvoller und gegebenenfalls sogar illegal – etwa im Fall von Schadprogrammen, die im Darknet zum Kauf angeboten werden. Hinzukommt, dass Angreifer oft auf größere finanzielle Ressourcen zugreifen als sie den betroffenen Institutionen oder KMU zur Abwehr zur Verfügung stehen.

2.3 Zielgruppenspezifische Herausforderungen

Kleine und mittlere Unternehmen: Bewusstsein und Hemmnisse

Beim Aufbau von Kompetenzen im Bereich der IT-Sicherheit müssen insbesondere die spezifischen Herausforderungen für kleine und mittlere Unternehmen (KMU) beachtet werden. Zwar ist das Bewusstsein für die Relevanz der IT-Sicherheit in KMU vorhanden: In einer Umfrage des Wissenschaftlichen Instituts für Infrastruktur und Kommunikationsdienste gaben zwei Drittel der befragten Unternehmen an, dass IT-Sicherheit eine hohe Bedeutung für sie habe. Jedoch haben nur 20 Prozent bereits eine Sicherheitsanalyse durchgeführt. Die Gründe für unzureichende technische und organisatorische IT-Sicherheitsmaßnahmen sind vielfältig: zu hoher Zeit- und Kostenaufwand, Mangel an entsprechend qualifiziertem Personal, aber auch die Unübersichtlichkeit der bestehenden Informations- und Beratungsangebote (Hillebrand et. al. 2017: 3, 11).

Bürgerinnen und Bürger: Beispiel Smart Home

IT-Sicherheit ist nicht nur für Unternehmen und Behörden ein entscheidendes Element für eine erfolgreiche Digitalisierung. Auch für Bürgerinnen und Bürger ergeben sich neue Herausforderungen. Dem Beratungsunternehmen Gartner zufolge wurden im Jahr 2017 etwa 63 Prozent der weltweit 8,4 Milliarden vernetzten Geräte von Konsumenten genutzt (Gartner 2018). Mehr und mehr dieser Geräte besitzen KI-Komponenten wie etwa lernende Thermostate und Heizungssysteme. Jedoch sind viele Geräte, die beispielsweise im Smart-Home-Umfeld verwendet werden, nicht ausreichend auf die Gewährleistung von Datensicherheit ausgerichtet und gegen Cyber-Angriffe geschützt. Sie könnten etwa von Angreifern übernommen und für Angriffe auf Dritte oder zum Eingriff in die Privatsphäre der Konsumenten missbraucht werden. Die systemeigenen KI-Komponenten könnten ebenfalls zum Schaden der Nutzer manipuliert oder ausgenutzt werden, etwa analog zu Crypto-Mining und Ransomware (siehe Kasten). Schließlich könnten KI-unterstützte Cyberangriffe auf Smart-Home-Geräte zukünftig deren Sicherheitsmechanismen leichter überwinden.

Diesen Herausforderungen und Risiken stehen Nutzer oft hilflos gegenüber: Sie erhalten beim Kauf derzeit weder eine Orientierung darüber, wie sicher die KI-Komponenten der Produkte sind, noch wie lange der Hersteller den vorhandenen Sicherheitsgrad durch Updates und Wartung aufrechterhalten wird. Das Bewusstsein bei Herstellern und Verbrauchern für die mit der Vernetzung verbundenen Risiken wächst jedoch (BSI 2018: 20). Die Politik hat entsprechende Kennzeichnungen bereits auf die Agenda gesetzt. In der Cyber-Sicherheitsstrategie der Bundesregierung werden wirksame und bedarfsgerechte Zertifikate und Gütesiegel als Instrument für die Verbreitung von Sicherheitsstandards genannt (Bundesministerium des Innern 2016: 17) und auch gemeinsam mit der Wirtschaft sollen IT-Sicherheitsstandards für internetfähige Produkte entwickelt werden (Koalitionsvertrag 2017: 45).

Crypto-Mining und Ransomware
Beim Crypto-Mining werden Endgeräte wie PCs und Smartphones von Schadsoftware infiziert, die deren Rechenleistung für das heimliche Erzeugen z. B. von Bitcoins einsetzt. Die Gerätebesitzer werden um Stromkosten, Akkulaufzeit und Geräteleistung betrogen. Analog wäre das illegale Profitieren der Angreifer von trainings- oder KI-gelernten Modellen der Endgeräte denkbar. Ähnlich wie bei Ransomware, wo Nutzerdaten kryptografisch gesperrt (verschlüsselt) und nur nach Zahlung von Lösegeld wieder entschlüsselt werden, könnte das Verhalten von KI-gesteuerten Systemen so manipuliert werden, dass ihre Besitzer für deren Wiederherstellung Lösegeld zahlen.

Kundendienstleistungen: Beispiel Chatbots

Für Kundendienstleistungen lassen sich in Unternehmen und Behörden künftig Chatbots einsetzen, um niedrigschwellig und passgenau Informationen und Leistungen bereitzustellen. Chatbots, die auf KI-Technologien zur Verarbeitung natürlicher Sprache und Mensch-Maschine-Interaktion basieren, könnten beispielsweise Anfragen zu Produkten, Dienstleistungen, Zuständigkeiten, Öffnungszeiten und Ansprechpartnern beantworten. Die Qualität der Auskunft und deren Personalisierung kann verbessert werden, indem KI-Systeme bereits während des Gesprächs den Dialog und die Daten der Nutzer auswerten und das Verhalten der Chatbots entsprechend optimieren. Die Ausgabe der Ergebnisse muss dabei nicht in Textform erfolgen. Es können auch aufgezeichnete Nachrichten oder Sprachsynthese zum Einsatz kommen. In diesem Bereich wurden mithilfe maschineller Lernverfahren bereits große Fortschritte erzielt. Zukünftig könnte es schwieriger werden, die Interaktion mit einem Menschen von der mit einem Bot eindeutig zu unterscheiden. Dabei muss auch die Sicherheit der Systeme beachtet werden. Cyberangriffe auf Chatbots – mit KI-Unterstützung oder auf die KI-Komponenten der Chatbots – können zwei Ziele verfolgen: Die Angreifer versuchen in den Besitz von (personenbezogenen) Daten zu kommen. Oder sie versuchen das Verhalten der Chatbots zu beeinflussen, etwa um deren Nutzer zu manipulieren und zur Herausgabe sensibler Informationen zu verleiten.

3. Herausforderung: Dual-Use-Potenzial von KI

Wie die meisten Technologien haben auch Methoden und Verfahren der Künstlichen Intelligenz einen Dual-Use-Charakter. Im Bereich der IT-Sicherheit können KI-Systeme oder maschinelle Lernverfahren zweckentfremdet und zu böswilligen oder kriminellen Zwecken genutzt werden. Das Bedrohungsrisiko durch eine Zweckentfremdung von KI-Systemen sollte nicht überzeichnet werden, jedoch sollten Entwickler und Anwender ein entsprechendes Bewusstsein für das Dual-Use-Potenzial entwickeln.

Eine viel beachtete Studie identifiziert drei mögliche zukünftige Effekte des Einsatzes von KI auf die Entwicklung von Cyberbedrohungen (Brundage et. al. 2018: 18f.):

- Erweiterung bestehender Angriffsstrategien, z. B. automatisierte Recherche von Informationen zur besseren Personalisierung von Spear-Phishing-Attacken.
- Entstehung neuer Bedrohungen, z. B. gezielte Angriffe auf die spezifischen Schwachstellen von KI-Systemen (siehe Kapitel 4).
- Veränderter Charakter der Attacken: Cyberattacken werden nicht nur effektiver, sondern auch effizienter und skalierbar.

Bei der böswilligen oder kriminellen Nutzung von KI-Technologien können sowohl Systeme zum Einsatz kommen, die für diese Art der Nutzung zweckentfremdet werden, als auch solche, die speziell für diese Nutzung entwickelt wurden.

Spear-Phishing-Attacken
Bei Spear-Phishing-Attacken versuchen Angreifer sensible Informationen zu erhalten oder auch ihre Ziele auf Webseiten zu lotsen, die Malware enthalten. Spear-Phishing-Attacken sind dabei auf ihre Opfer (Personen, Organisationen oder Unternehmen) zugeschnitten, um die Erfolgswahrscheinlichkeit zu erhöhen. Die dafür notwendige Identifikation der Ziele sowie deren sozialer und professioneller Netzwerke ist dabei deutlich aufwändiger als bei herkömmlichen Phishing-Versuchen.

Szenarien für den böswilligen oder kriminellen Einsatz von KI-Technologien

Social Engineering mit KI-Unterstützung

Social-Engineering-Methoden zielen darauf ab, die Nutzer – seien es Privatpersonen oder Beschäftigte eines Unternehmens – dazu zu verleiten, selbstständig Daten preiszugeben oder Malware auf ihren Systemen zu installieren. Dabei werden auch menschliche Schwächen wie Neugier oder Angst ausgenutzt (BSI 2018: 98). Den Einsatz von Social-Engineering-Methoden begünstigen einerseits die zunehmenden Fähigkeiten der Systeme im

Bereich sozialer Interaktion (siehe Einsatz von Chatbots, Kapitel 2). Zugleich lassen sich mithilfe von KI-Systemen Teile des Social-Engineering-Prozesses automatisieren, indem gezielt online verfügbare Informationen extrahiert und dazu genutzt werden, betrügerische Websites, Links oder Mails zu generieren, die auf ihre Zielperson zugeschnitten sind (Brundage et. al 2018.: 24).

■ KI-unterstützte Analyse und Tests von Schwachstellen

Die Fähigkeiten von Schadprogrammen, IT-Systeme in Unternehmen, Behörden oder im Smart Home automatisiert auf Schwachstellen zu überprüfen, lassen sich durch den Einsatz von KI erhöhen. Dabei können beispielsweise bekannte Muster von Code-Schwachstellen gelernt und genutzt werden, um die Entdeckung neuer Schwachstellen zu beschleunigen (Brundage et. al. 2018: 24). Hier könnte auch KI-verbessertes Fuzzing zum Einsatz kommen.

Fuzzing

Fuzzing ist eine Technik, um Programmierfehler und Sicherheitslücken zu finden. Dabei wird ein Computerprogramm mit vielen zufälligen, eventuell sinnlosen Eingaben getestet. Die Hauptschwierigkeit ist die Wahl der Eingaben, die zufällig sein, aber dennoch möglichst viele Stellen der Software erreichen sollte. Lernende Systeme, die die Eingaben variieren und dabei die Anzahl der erreichten Stellen in der Software optimieren, könnten hier ein vielversprechender Ansatz für Angreifer sein.

Bei der automatisierten Analyse sowie bei Tests von Schwachstellen können Angreifer das Zeitfenster zwischen Veröffentlichung einer Schwachstelle und deren Schließung ausnutzen. Gemeldete Schwachstellen werden mit einer Kennung, der sogenannten CVE-Nummer (Common Vulnerabilities and Exposures), versehen und mit einer Beschreibung veröffentlicht. Ein entsprechend trainierter Algorithmus könnte diese Informationen automatisiert auslesen und anschließend Websites auf die Existenz dieser Schwachstelle prüfen.

■ KI-unterstützte Optimierung von Schadsoftware und Angriffen

Ebenso wie Systeme zur Angriffserkennung auf Basis bekannter Angriffsmuster trainiert werden, könnten Angreifer aus dem Verhalten der Sicherheitssysteme bei Angriffen lernen und ihre Schadprogramme entsprechend optimieren. Da die Sicherheitstechnologien am Markt frei verfügbar sind, könnten Angreifer ihre Systeme im Labor gegen die Sicherheitssysteme antreten lassen und daraus Erkenntnisse für die Optimierung gewinnen. So ließe sich beispielsweise ein Modell erlernen, welche Netzverkehre von einem Angriffserkennungssystem als legitim betrachtet werden. Auf dieser Grundlage simuliert die Schadsoftware legitime Netzverkehre, um die schadhaften Aktivitäten zu verdecken. So lassen sich beispielsweise Distributed-denial-of-service-Attacken verbessern, die darauf abzielen, einen Dienst durch eine große Masse von Anfragen zu überlasten. Imitieren die angreifenden Systeme (in der Regel Rechner aus einem Bot-Netz) menschliches Verhalten (z. B. typische Klick- und Navigationsmuster auf einer angegriffenen Website), wird ein solcher Angriff möglicherweise nicht oder zu spät erkannt.

Umgehung von Authentisierungsverfahren

Zukünftig werden bild- und sprachbasierte Authentisierungsverfahren durch die Möglichkeiten herausgefordert, mithilfe von KI-Systemen Bilder und Videos zu manipulieren oder Stimmen zu synthetisieren und zu imitieren. Systeme zur Anwesenheitserkennung, die anhand von Merkmalen wie Tippmustern beurteilen, ob tatsächlich eine Verbindung zu einem legitimen Nutzer besteht, könnten durch KI-Systeme unterlaufen werden, die das Verhalten der Nutzer auf Basis eines zuvor trainierten Modells imitieren. Diese Nachahmung legitimen Nutzerverhaltens und als legitim betrachteter Netzverkehre, gepaart mit zuvor gestohlenen Anmeldeinformationen, kann genutzt werden, um Zugriff auf geheime Informationen von Unternehmen oder Behörden zu erhalten.

Für die hier beschriebenen Szenarien könnten oftmals auch KI-Werkzeuge genutzt werden, die ursprünglich für andere, nicht kriminelle Zwecke entwickelt wurden. KI-Systeme, die speziell für den bösartigen oder kriminellen Einsatz entwickelt werden, sind eine eigenständige Herausforderung für die Gefahrenabwehr und derzeit vermutlich noch staatlichen Akteuren oder kriminellen Organisationen mit entsprechenden Ressourcen vorbehalten. Denn Angriffe mithilfe von KI-Systemen oder maschinellen Lernverfahren sind durchaus voraussetzungsvoll. Benötigt wird einerseits die richtige Infrastruktur wie die Rechenleistung und entsprechende (große) Datensätze zum Training der Systeme. Möglich wäre, dass Angreifer die notwendigen Rechenkapazitäten von Hosts und Rechenzentren abzweigen, die vorher mittels Malware infiziert wurden. Andererseits erfordert auch die Entwicklung leistungsfähiger Algorithmen entsprechendes Wissen sowie zeitliche und finanzielle Ressourcen. Der Zugang zu entsprechender Software und relevanten wissenschaftlichen Erkenntnissen ist jedoch relativ einfach. Viele neue KI-Algorithmen werden innerhalb von Wochen und manchmal Tagen reproduziert und viele wissenschaftliche Publikationen liefern den Quellcode gleich mit (Brundage et. al. 2018: 17).

Bezogen auf den flächendeckenden Einsatz von KI-Systemen zur Durchführung oder Unterstützung von Cyber-Angriffen kann zum gegenwärtigen Zeitpunkt jedoch festgehalten werden, dass Angreifer noch im Hintertreffen sind und die seriösen Entwickler einen entsprechenden Vorsprung besitzen.

4. Sicherheit und Schutz von KI-Systemen

Lernende Systeme werden zunehmend in Prozesse integriert, bei denen Sicherheit und Datenschutz eine zentrale Rolle spielen. So übernehmen beispielsweise Algorithmen des maschinellen Lernens wichtige Funktionen in autonomen Fahrzeugen, dem Wertpapierhandel und der digitalen Gesundheitsversorgung. In diesen Anwendungsfeldern müssen die Systeme nicht nur zuverlässige Vorhersagen treffen, sondern auch Angriffen und Manipulationen widerstehen. Sicherheitsprobleme, die erst durch den Einsatz von KI entstehen, müssen frühzeitig erkannt und behoben werden.

Viele Algorithmen des maschinellen Lernens – insbesondere neuronale Netze – sind jedoch nicht für eine Anwendung in unsicheren Umgebungen ausgelegt und können durch verfälschte Eingaben während des Lernens oder der Vorhersage gestört werden. So ist es beispielsweise möglich, Lernende Systeme in Fahrzeugen zu irritieren (Eykholt et. al. 2018) oder durch geschickt formulierte Anfragen vertrauliche Daten aus medizinischen Systemen zu extrahieren, wenn die zugrunde liegenden Algorithmen unzureichend geschützt sind (Fredrikson et. al. 2018). Es ist zu erwarten, dass derartige Angriffe mit dem wachsenden Einsatz von KI an Bedeutung gewinnen.

Erst seit wenigen Jahren setzt sich die Forschung mit dieser Problematik auseinander und ein systematisches Verständnis von Sicherheit und Datenschutz in Lernenden Systemen ist erst in der Entstehung. Grundsätzlich werden in der Forschung drei Arten von Angriffen auf Lernende Systeme unterschieden, die unterschiedliche Phasen des Lernvorgangs stören.

4.1 Angriffe gegen Lernalgorithmen

■ Angriffe auf die Vorhersage

Lernende Systeme nutzen verschiedene Datenquellen für die Erkennung von Objekten und die Vorhersage von Ereignissen (z. B. Sensordaten in autonomen Fahrzeugen). Nicht immer kann die Integrität dieser Daten sichergestellt werden, sodass manipulierte Eingaben das System täuschen und zu Fehlentscheidungen führen können. Angreifer können hierbei insbesondere blinde Flecken in den gelernten Modellen ausnutzen, die schon bei geringer Manipulation zu falschen Vorhersagen führen und durch Menschen kaum nachvollziehbar sind. Ein bekanntes Beispiel ist die Manipulation eines Stoppschilds mithilfe von Aufklebern. Das Lernende System in einem Fahrzeug konnte daraufhin das Schild nicht mehr richtig als Stoppschild erkennen (Eykholt et. al. 2018).

■ Angriffe während des Lernens

Für das Anlernen eines KI-Systems sind oftmals umfangreiche Trainingsdaten notwendig, die Millionen von Beispielen umfassen können. Die Integrität derartiger Datensätze lässt sich in der Praxis schwer verifizieren und so entsteht die Möglichkeit, dass manipulierte Beispiele in den Trainingsdaten die Funktionsweise eines Lernenden Systems beeinträchtigen. Die Auswirkungen hängen hierbei von der Stärke des Angriffs ab und können von einer reduzierten Vorhersagegenauigkeit bis zu gezielten Fehlentscheidungen führen. Ein eindrückliches Beispiel hierfür sind sogenannte Hintertüren in neuronalen Netzen, bei denen der Angreifer Fehlentscheidungen vorprogrammiert, beispielsweise falsche Lenkbewegungen in speziellen Situationen (Liu et. al 2018).

■ Angriffe auf den Datenschutz

Die dritte Klasse von Angriffen zielt auf die Extraktion von Informationen aus Lernenden Systemen. In vielen Anwendungsfeldern der KI, so zum Beispiel in der digitalen Gesundheitsversorgung, werden äußerst sensible Daten für den Lernvorgang eingesetzt. Bei fehlendem Schutz können einzelne Merkmale dieser Daten durch geschickte Eingaben aus dem System extrahiert und Personen zugeordnet werden. Ebenso existieren Angriffsstrategien, die es ermöglichen, gelernte Modelle aus einem System durch manipulierte Eingaben zu rekonstruieren. Ein Beispiel sind Angriffe auf Lernende Systeme zur Medikamentendosierung, bei denen genetische Merkmale von Patienten abgeleitet werden können (Fredrikson et. al. 2018).

4.2 Schutz- und Abwehrmaßnahmen

Der zukünftige Einsatz von Lernenden Systemen in sicherheitskritischen Anwendungen verlangt daher besondere Sorgfalt und die Integration von Schutz- und Abwehrmaßnahmen. Zum einen müssen grundsätzliche Prinzipien der Sicherheit Beachtung finden, beispielsweise der Schutz von Daten und Kommunikation durch moderne Verschlüsselungsverfahren. Zum anderen müssen neuartige Schutzkonzepte vorangetrieben werden, die speziell Lernalgorithmen schützen und langfristig einen sicheren Betrieb von KI ermöglichen. Vier Forschungsrichtungen sind hierbei von Bedeutung.

■ Resilienz gegen Manipulationen

Ein Zweig der Forschung beschäftigt sich mit der Härtung von Lernenden Systemen gegen manipulierte Eingaben. Da sich auch Menschen durch verfälschte Daten täuschen lassen, muss die Forschung hier zunächst versuchen, zur Sicherheit von menschlichen Entscheidungen aufzuschließen. Hierfür gibt es bereits zahlreiche Ansätze, die die Verarbeitung von Daten und die Ableitung von Vorhersagen in KI-Systemen mit unterschiedlichen technischen Konzepten gegen Manipulationen schützen. Die Wirksamkeit dieser Ansätze wird in der wissenschaftlichen Gemeinschaft zurzeit erprobt und es besteht ein klarer Bedarf an weiterer Intensivierung dieser Forschung (Carlini & Wagner 2018).

Testverfahren für KI

Eine weitere Strategie ist das sicherheitsorientierte Testen der angestrebten Funktionalität. Wie in anderen Bereichen der Produktentwicklung schon länger üblich, müssen auch Lernalgorithmen gegen ein breites Spektrum von Testfällen evaluiert werden. Die Forschung muss hierfür Techniken entwickeln, um gezielt ungewöhnliche Eingaben und somit potenzielle Angriffe automatisch zu generieren. Es soll so möglich werden, nicht nur wie bisher das Verhalten im Regelfall zu verifizieren, sondern gerade auch Sonderfälle zu untersuchen und Sicherheitsprobleme aufzudecken (Pei et. al. 2017).

Datenschutzerhaltendes Lernen

Ein weiterer Forschungszweig behandelt den Schutz von Daten in Lernenden Systemen. In den vergangenen Jahren sind Lernalgorithmen entstanden, die die Verknüpfungen zwischen Trainingsdaten und Vorhersagen durch technische Maßnahmen erschweren und einer Rekonstruktion von persönlichen Daten entgegenwirken. Es konnten so erste datenschutzerhaltende Lernalgorithmen entwickelt werden, die beispielsweise auch ein verteiltes Lernen ohne Austausch von Trainingsdaten ermöglichen (Shokri & Shmatikov 2018).

Erklärbarkeit von KI

Viele Lernalgorithmen verfügen über eine eingeschränkte oder sogar keine Möglichkeit zur Erklärung ihrer Entscheidungen. Im Bereich der Sicherheit sind solche Black-Box-Ansätze problematisch und es müssen Anstrengungen unternommen werden, um die Entscheidungen von Lernenden Systemen so gut wie möglich zu erklären. Erste Arbeiten aus der Forschung zeigen, dass auch komplexe neuronale Netze zu einem gewissen Grad Möglichkeiten zur Erklärung bieten. Diese Forschung muss weiter vorangetrieben werden, um den Einsatz von KI langfristig transparenter und sicherer zu gestalten.

5. Lösungsansätze

Der vermehrte Einsatz von KI-Systemen bringt enorme Potenziale für den Digitalstandort Europa und erzeugt im Feld der IT-Sicherheit eine neue Dynamik. Gleichzeitig ist die Gewährleistung von IT-Sicherheit im KI-Kontext ein Ausdruck der digitalen Souveränität. Ein gewinnbringender Einsatz von KI-Systemen zur Verbesserung der IT-Sicherheit, das Dual- Use-Potenzial von KI-Systemen sowie der Schutz der KI-Systeme selbst implizieren dabei unterschiedliche Handlungsfelder für Unternehmen, Politik und Behörden und die Forschung. Die nachfolgend skizzierten Lösungsansätze geben eine erste Einschätzung der Autoren dieses Papiers wieder und werden innerhalb der Plattform Lernende Systeme weiterentwickelt.

Allgemeine Handlungsfelder

- Der zukünftige Einsatz von Lernenden Systemen in sicherheitskritischen Anwendungen verlangt möglicherweise besondere Sorgfalt und die **Integration von spezifischen Schutz- und Abwehrmaßnahmen**. Dabei sollten grundsätzliche Prinzipien der Sicherheit beachtet und neuartige Verteidigungskonzepte vorangetrieben werden, die speziell Lernalgorithmen schützen und langfristig einen sicheren Betrieb von KI ermöglichen. Die Methodik der beweisbaren Sicherheit könnte dabei auch im Zusammenhang mit KI-Systemen angewendet werden.
- KI-Systeme sollten **mit einer technischen Rückfallebene ausgestattet** werden – für den Fall von Fehlfunktionen, Angriffen auf das System oder falls das System selbst sicherheitskritisches Verhalten zeigt. Der sichere Betrieb des Gesamtsystems darf dadurch nicht gefährdet werden.
- Mit Blick auf das Dual-Use-Potenzial empfiehlt es sich, unterschiedliche Möglichkeiten zu prüfen, wie die **Zweckentfremdung von KI-Systemen bestmöglich verhindert** werden kann. Dafür könnten zunächst gesellschaftliche und rechtliche Maßstäbe entwickelt werden.
- **Im Bereich von Spezialanwendungen** wie der Seitenkanalanalyse erscheint es eher unwahrscheinlich, dass in absehbarer Zeit ausgereifte Produkte zur Verfügung stehen. Vielmehr sollten Nutzer solcher Anwendungen (z. B. Wirtschaft, Universitäten und Behörden) bei Bedarf **eigene Expertise aufbauen**.
- Die heutige Fokussierung auf Betreiber kritischer Infrastrukturen sollte mit Blick auf vernetzte Geräte mit KI-Komponenten, die bei Bürgerinnen und Bürgern oder Behörden im Einsatz sind, ergänzt werden. Dabei sollten auch in Forschung und Entwicklung von Systemen mit KI-Komponenten die Prinzipien **Security by Design** und **Security by Default** beachtet werden.
- Angesichts der notwendigen Vertrauenswürdigkeit von KI-Systemen, u. a. auch in sicherheitskritischen Kontexten, ist eine forschungs- und industriepolitische Pointierung des Anspruchs der **Digitalen Souveränität**[2] notwendig.

2 Die Digitale Souveränität umfasst „die vollständige Kontrolle über gespeicherte und verarbeitete Daten sowie die unabhängige Entscheidung darüber, wer darauf zugreifen darf. Sie umfasst weiterhin die Fähigkeit, technologische Komponenten und Systeme eigenständig zu entwickeln, zu verändern, zu kontrollieren und durch andere Komponenten zu ergänzen. Digitale Souveränität ist deswegen einerseits wichtige Grundlage für vertrauenswürdige Systeme und andererseits unverzichtbare Voraussetzung für unabhängiges staatliches Handeln." (Plattform Innovative Digitalisierung der Wirtschaft 2018: 3.)

Handlungsfelder für Politik und Behörden

- Um die Fachkräftelücke im Bereich der IT-Sicherheit zu schließen, sind die Anstrengungen im Bereich der **Ausbildung und Gewinnung von Fachkräften in der IT-Sicherheit** zu verstärken. Dabei könnten der Umgang mit KI-Systemen für die IT-Sicherheit in der Aus- und Weiterbildung der Fachkräfte berücksichtigt und entsprechend Lehrpläne auf dem Stand der Technik fortlaufend aktualisiert werden.
- Ein **Grundverständnis für IT-Sicherheit** sollte in angemessener Form auch in Disziplinen integriert werden, in denen das Thema im Zuge der fortschreitenden Digitalisierung zunehmend an Bedeutung gewinnt, etwa im Maschinenbau.
- Für KMU empfiehlt es sich, beispielsweise im Rahmen der bestehenden Kompetenzzentren Angebote zu schaffen oder auszubauen, mit deren Hilfe sie ihre **Kompetenzen im Bereich der IT-Sicherheit im Hinblick auf den Einsatz von KI-Systemen erweitern** können. Dabei sollten auch entsprechende Beratungsangebote berücksichtigt werden. Ein Navigationssystem zu IT-Sicherheit im KI-Kontext könnte KMU unterstützen und dazu beitragen, bestehende Angebote übersichtlicher und zugänglicher zu gestalten.
- Bei der **Integration von KI-Systemen in die öffentliche Verwaltung** und in Dienste für Bürgerinnen und Bürger spielt die Sicherheit der Systeme eine wichtige Rolle. Setzen Behörden beispielsweise Chatbots für ihre Services ein, sollten Maßnahmen erarbeitet und ergriffen werden, die verhindern, dass Angreifer durch Überlisten der für diese Dienste eingesetzten KI-Komponenten Zugriff auf personenbezogene Daten erhalten.
- Der Einsatz von KI-Systemen verdeutlicht den Bedarf einer **kohärenten, möglichst globalen IT-Sicherheitspolitik**. Internationale Initiativen sollten im Ergebnis auch auf globaler Ebene verantwortliches staatliches Handeln in diesem Bereich fördern, etwa ein entschlossenes Vorgehen gegen Hackerangriffe vom eigenen Territorium.

Handlungsfelder für Unternehmen

- Konzerne sollten Werkzeuge wie **Angriffserkennungssysteme mit KI-Funktionalität** am Markt kaufen und selbst betreiben, während kleinere Firmen Dienstleistungen beziehen könnten. Wichtig ist, dass entsprechende Angebote für KMU gemacht werden.
- Der Aufbau technischer Fähigkeiten und Kompetenzen für den Umgang mit KI im Bereich der IT-Sicherheit ist möglicherweise erfolgskritisch. Marktfähige **KI-unterstützte Sicherheitslösungen und deren ständige Weiterentwicklung** sind wichtige Voraussetzungen für die IT-Sicherheit der deutschen Industrie.
- Unternehmen sollten ihre **Maßnahmen und Kompetenzen in der IT-Sicherheit** mit Blick auf den zukünftigen Einsatz von KI in diesem Bereich überprüfen und gegebenenfalls Anstrengungen unternehmen, entsprechende Kompetenzen aufzubauen.

- In Verbindung mit KI können künftige Angriffe auf die IT-Systeme in Büros und in der Produktion gezielter und deutlich intelligenter ausgeübt werden. Unternehmen könnten deshalb neben den klassischen Schwachstellen- und Bedrohungsanalysen **KI-unterstützte und lernende Überwachungssysteme** implementieren.
- Wie auch in der IT-Sicherheit schreitet der Stand der Technik auch bei KI-Systemen voran. Erforderlich ist daher eine revolvierende **Überprüfung der bereits eingesetzten intelligenten Abwehrmaßnahmen** inklusive der verwendeten KI.

Handlungsfelder im Bereich Forschung

- Es besteht Forschungsbedarf, wie sich KI-Systeme für die Unterstützung und Verbesserung der IT-Sicherheit nutzen lassen, aber auch mit Blick auf die Sicherheit und den Schutz der KI-Systeme selbst. Dazu sollten entsprechende Forschungsaktivitäten in diesem Bereich angestoßen oder ausgebaut werden. Sie könnten unterstützt werden durch eine **stärkere Vernetzung von Einrichtungen mit Forschungsschwerpunkten** im KI-Bereich und/oder im Bereich der IT-Sicherheit in Verbindung mit einem weiteren Kompetenzaufbau in Deutschland.
- Die Möglichkeiten, die der Einsatz von KI-Systemen für Bürgerinnen und Bürger, Unternehmen oder die öffentliche Verwaltung in unterschiedlichen Anwendungsbereichen bietet, lassen sich nur nutzen, wenn die Systeme möglichst gut gegen Manipulationen geschützt sind, insbesondere gegen Angriffe auf die Vorhersage oder Angriffe auf den Lernprozess. Hier lässt eine Intensivierung der Forschung zur **Resilienz von KI-Systemen gegen Manipulationen** entscheidende Fortschritte erwarten.
- Die Sicherheit von KI-Systemen lässt sich insbesondere durch **Tests gegen Sonderfälle** erhöhen. Die Forschung zu Techniken, die beispielsweise gezielt ungewöhnliche Eingaben automatisiert generieren und damit potenzielle Angriffe simulieren, könnte hier einen wichtigen Beitrag leisten.
- Auch beim Einsatz von KI-Systemen muss der Schutz personenbezogener Daten sichergestellt werden, insbesondere in Anwendungsfeldern, in denen sensible Daten für den Lernprozess der Systeme verwendet werden, wie beispielsweise im medizinischen Bereich. Die fortlaufende Erforschung und **Weiterentwicklung datenschutzerhaltender Lernalgorithmen**, die eine Extraktion und Rekonstruktion personenbezogener Daten aus den Modellen der Lernenden Systeme verhindern oder erschweren, kann dazu beitragen, den Einsatz von KI-Systemen in unterschiedlichen Anwendungsbereichen unter Wahrung des Datenschutzes zu befördern.
- In der IT-Sicherheit, wie in anderen Anwendungsfeldern von KI-Systemen, kann die **Erklärbarkeit der Entscheidungen der Systeme** zu einem wichtigen Faktor für deren Anwendbarkeit werden. Das gilt etwa in Anwendungsbereichen, in denen der Mensch, der mit den Systemen interagiert, die Einflussfaktoren für deren Entscheidungen nachvollziehen und beurteilen können muss. Dies betrifft insbesondere komplexe neuronale Netze. Die Forschung zu Möglichkeiten der Erklärbarkeit sollte vorangetrieben werden, um einen sicheren und transparenten Einsatz der Systeme zu ermöglichen.

Autoren und Redaktion

Autorinnen und Autoren
Prof. Dr. Jörn Müller-Quade, Karlsruher Institut für Technologie
Dr. Gisela Meister, Giesecke + Devrient Mobile Security GmbH
Prof. Dr. Thorsten Holz, Ruhr-Universität Bochum
Dr. Detlef Houdeau, Infineon Technologies AG
Prof. Dr. Konrad Rieck, Technische Universität Braunschweig
Peter Rost, Rohde & Schwarz Cybersecurity GmbH
Thomas Schauf, Deutsche Telekom AG
Prof. Dr. Werner Schindler, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Die Autorinnen und Autoren sind Mitglieder der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme.

Redaktion
Johannes Melzer, Geschäftsstelle Plattform Lernende Systeme

Über die Plattform Lernende Systeme

Lernende Systeme im Sinne der Gesellschaft zu gestalten – mit diesem Anspruch wurde die Plattform Lernende Systeme im Jahr 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert. Die Plattform bündelt die vorhandene Expertise im Bereich Künstliche Intelligenz und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter. Die rund 200 Mitglieder der Plattform sind in Arbeitsgruppen und einem Lenkungskreis organisiert. Sie zeigen den persönlichen, gesellschaftlichen und wirtschaftlichen Nutzen von Lernenden Systemen auf und benennen Herausforderungen und Gestaltungsoptionen.

Literatur

Bundesamt für Sicherheit in der Informationstechnik (Hg.) (2018): Die Lage der IT-Sicherheit in Deutschland 2018. Bonn.

Bitkom (2017): Wirtschaftsschutz in der digitalen Welt. www.bitkom.org/Presse/Anhaenge-an-PIs/2017/07-Juli/Bitkom-Charts-Wirtschafts-schutz-in-der-digitalen-Welt-21-07-2017.pdf (abgerufen am 18.03.2019).

Brundage et. al. (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf (abgerufen am 18.03.2019).

Bundesministerium des Innern (2016): Cyber-Sicherheitsstrategie für Deutschland 2016. www.bmi.bund.de/cybersicherheitsstrategie/BMI_CyberSicherheitsStrategie.pdf (abgerufen am 18.03.2019)

Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung. www.bmbf.de/files/Nationale_KI-Strategie.pdf (abgerufen am 18.03.2019).

Capgemini Digital Transformation Institute (Hg.) (2018): Cybersecurity Talent. The Big Gap in Cyber Protection. Eight Recommendations for How Organizations Can Bridge the Cybersecurity Talent Gap. www.capgemini.com/de-de/wp-content/uploads/sites/5/2018/02/the-cybersecurity-talent-gap-v8_web-2.pdf (abgerufen am 18.03.2019).

Carlini/Wagner (2017): Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP). https://arxiv.org/pdf/1608.04644.pdf (abgerufen am 18.03.2019).

Eykholt et al. (2018): Robust Physical-World Attacks on Deep Learning Models. CVPR, 2018. https://arxiv.org/pdf/1707.08945.pdf (abgerufen am 18.03.2019).

Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e. V. (Hg.) (2018): Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung. München.

Fredrikson et. al. (2014): Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proc USENIX Secur Symp. 2014. www.ncbi.nlm.nih.gov/pmc/articles/PMC4827719/ (abgerufen am 18.03.2019).

Gartner (2017): Gartner Says 8.4 Billion Connected „Things“ Will Be in Use in 2017, Up 31 Percent From 2016. www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016 (abgerufen am 18.03.2019).

Gógl (2018): Neue Methoden der Authentifizierung.
www.cio.de/a/neue-methoden-der-authentifizierung,3581563 (abgerufen am 18.03.2019)

Hillebrand et al. (2017): Aktuelle Lage der IT-Sicherheit in KMU. Kurzfassung der Ergebnisse der Repräsentativbefragung.
www.wik.org/fileadmin/Sonstige_Dateien/IT-Sicherheit_in_KMU/Aktuelle_Lage_der_IT-Sicherheit_in_KMU_-_WIK.pdf (abgerufen am 18.03.2019).

IDG Research Services (2018): Studie Machine Learning / Deep Learning 2018.
www.lufthansa-industry-solutions.com/de-de/studien/idg-studie-machine-learning-2018/?gclid=EAIaIQobChMItI6skLW13gIVy7ftCh28swBiEAAYASAAEgKUlvD_BwE (abgerufen am 18.03.2019).

Karapanos (2018): Cybersicherheit und Künstliche Intelligenz.
www.welove.ai/de/blog/post/cybersicherheit-und-kuenstliche-intelligenz.html (abgerufen am 18.03.2019)

Koalitionsvertrag (2017): Ein neuer Aufbruch für Europa. Eine neue Dynamik für Deutschland. Ein neuer Zusammenhalt für unser Land. Koalitionsvertrag zwischen CDU, CSU und SPD.
www.cdu.de/system/tdf/media/dokumente/koalitionsvertrag_2018.pdf?file=1 (abgerufen am 18.03.2019)

Liu et al. (2018): Trojaning Attacks on Neural Networks. Network and Distributed Systems Security (NDSS) Symposium, 2018.
www.cs.purdue.edu/homes/ma229/papers/NDSS18.TNN.pdf (abgerufen am 18.03.2019)

Nern (2018): Cybersecurity-Trends: Das erwartet uns 2018.
www.ibm.com/de-de/blogs/think/2018/01/23/cybersecurity-trends-2018/ (abgerufen am 18.03.2019).

Pei et. al. (2017): DeepXplore: Automated Whitebox Testing of Deep Learning Systems. SOSP, 2017. https://arxiv.org/pdf/1705.06640.pdf (abgerufen am 18.03.2019).

Plattform Innovative Digitalisierung der Wirtschaft (2018): Digitale Souveränität und Künstliche Intelligenz – Voraussetzungen, Verantwortlichkeiten und Handlungsempfehlungen. www.de.digital/DIGITAL/Redaktion/DE/Digital-Gipfel/Download/2018/p2-digitale-souveraenitaet-und-kuenstliche-intelligenz.pdf?__blob=publicationFile&v=5 (abgerufen am 18.03.2019).

Plattform Industrie 4.0 (2019): Künstliche Intelligenz (KI) in Sicherheitsaspekten der Industrie 4.0.
https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-in-sicherheit-saspekten.pdf?__blob=publicationFile&v=4 (abgerufen am 02.04.2019)

Shokri/Shmatikov (2016): Privacy-Preserving Deep Learning. ACM CCS.
www.comp.nus.edu.sg/~reza/files/Shokri-CCS2015.pdf (abgerufen am 18.03.2019).

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
kontakt@plattform-lernende-systeme.de
www.plattform-lernende-systeme.de
Twitter: @LernendeSysteme

Gestaltung
PRpetuum GmbH, München

Stand
April 2019

Bildnachweis
matejmo / iStock / Titel

Dieses Werk ist urheberrechtlich geschützt.
Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, der Entnahme von Abbildungen, der Wiedergabe auf fotomechanischem oder ähnlichem Wege und der Speicherung in Datenverarbeitungsanlagen, bleiben – auch bei nur auszugsweiser Verwendung – vorbehalten.