Norbert Pohlmann

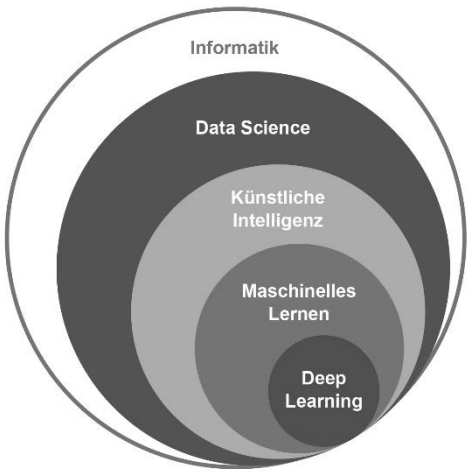
IT-Sicherheit und Künstliche Intelligenz

IT-Sicherheit und Künstliche Intelligenz (KI) sind zwei wichtige Technologien, die sich gegenseitig stark beeinflussen. Dieser Artikel beginnt mit den grundlegenden Einordnungen, Definitionen, Begriffen und Prinzipien rund um die Künstliche Intelligenz. Danach wird das Thema Künstliche Intelligenz für IT-Sicherheit behandelt. Hier geht es darum, wie die KI helfen kann, die IT-Sicherheit zu verbessern. Anschließend wird diskutiert, wie Angreifer die KI nutzen und wie sich das auf die IT-Sicherheit der Verteidiger auswirkt. Weiterhin beleuchtet der Artikel das wichtige Thema IT-Sicherheit für KI. Dabei wird dargestellt, was getan werden muss, damit die KI, die genutzt wird, nicht durch die Angreifer manipuliert werden kann.

1 Einordnung der Künstlichen Intelligenz

„Data Science" ist eine Wissenschaft des Fachgebiets der Informatik und beschäftigt sich mit der Extraktion von Wissen aus den Informationen in Daten (Abb. 1). Da es immer mehr Daten mit Informationen gibt, kann auch immer mehr Wissen aus den Informationen der Daten abgeleitet werden.

Abb. 1 | Einordnung der Künstlichen Intelligenz



Norbert Pohlmann

ist Informatikprofessor für Cyber-Sicherheit und Leiter des Instituts für Internet-Sicherheit - if(is) an der Westfälischen Hochschule in Gelsenkirchen sowie Vorstandsvorsitzender des Bundesverbands IT-Sicherheit – TeleTrusT und im Vorstand des Internetverbandes - eco.
E-Mail: pohlmann@internet-sicherheit.de

Bei Künstlichen Intelligenzen wird zwischen schwacher und starker KI unterschieden. Starke „Künstliche Intelligenz" soll automatisiert „menschenähnliche Intelligenz" nachbilden. Die Begriffe „Singularität" oder „Artificial General Intelligence (AGI)" beziehen sich auf einen hypothetischen Zeitpunkt, bei dem das KI-System eine höhere (künstliche) Intelligenz besitzt als die (menschliche) Intelligenz. KI-Systeme können sich dann selbstständig verbessern und eigenständig sehr schnell Fortschritte erzielen, die für die Menschheit nicht mehr vorhersehbar sind. Filme wie der Terminator oder Bücher wie Origin von Dan Brown zeigen sehr schön, dass wir Singularität eigentlich nicht haben wollen, weil die Gefahr besteht, dass sie sich verselbstständigt. Die Menschheit verliert die Kontrolle über die KI, die Zukunft wird unvorhersehbar und das kann negative Auswirkungen haben. Unsere gemeinsame Aufgabe muss es daher sein, sicherzustellen, dass KI-Systeme, welche die menschliche Intelligenz übertreffen, im Einklang mit menschlichen Werten und Zielen agieren.

Unter schwache „Künstliche Intelligenz" wird Maschinelles Lernen (ML) verstanden, was heute sehr erfolgreich ist und in letzter Zeit durch neue Innovationen einen sehr großen Hype ausgelöst hat. Maschinelles Lernen ist ein Begriff im Bereich der Künstlichen Intelligenz und steht für die „künstliche" Generierung von Wissen aus den Informationen in Daten mit der Hilfe von IT-Systemen [1].

Deep Learning ist eine Methode des maschinellen Lernens. Deep Learning war eine wichtige Innovation, weil dadurch eine effizientere Verarbeitung von komplexen Daten ermöglicht wurde und die Ergebnisse deutlich verbessert wurden. Die letzte große Innovation ist das Large Language Modell (LLM), als Basismodell für generative KI.

Beim maschinellen Lernen ist der Begriff „Stochastischer Papagei" eine Metapher zur Beschreibung der Theorie, dass ein großes Sprachmodell zwar in der Lage ist, plausibel Sprache zu erzeugen, deren Bedeutung jedoch nicht versteht. Daher kann ein großes Sprachmodell auch keinen Fehler identifizieren und der Mensch muss in der Lage sein die Ergebnisse selbst zu verifizieren, wenn er die großen Sprachmodelle verantwortungsvoll nutzen will.

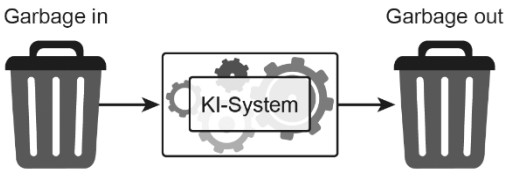
Generative KI (GenAI) ist eine Form der künstlichen Intelligenz, die gestützt auf ihren Trainingsdaten Texte, Bilder, Code

und verschiedene andere Inhalte produzieren kann. ChatGPT ist das bekannteste Beispiel und hat eine Revolution in der Digitalisierung auf sehr vielen verschiedenen Ebenen ausgelöst.

2 Paradigma: Garbage in, Garbage out (GIGO)

„Garbage in – Garbage out" ist ein Paradigma in der KI, das eine wichtige Rolle für die Qualität der Ergebnisse von KI-Systemen spielt (Abb. 2). Beim KI geht es um die „Extraktion von Wissen aus Daten" und das bedeutet, wenn in den verwendeten Daten keine Informationen enthalten sind, kann die KI daraus auch kein Wissen extrahieren. Das können wir uns sehr gut vorstellen, wenn alle Daten null sind. Generell gilt: Wenn die Qualität der Eingabedaten schlecht ist, wird auch die Qualität der Ergebnisse schlecht sein. Aus diesem Grund werden als Input hochqualitative Daten benötigt, damit mithilfe der KI vertrauenswürdige Ergebnisse erzielt werden können.

Abb. 2 | Garbage in, Garbage out



2.1 Qualität der Eingabedaten

Wegen der Wichtigkeit der Qualität der Eingabedaten sollte ein Standard der Datenqualität für KI-Systeme etabliert werden. Im Einzelnen sind dabei unter anderem Vollständigkeit, Repräsentativität, Nachvollziehbarkeit, Aktualität und Korrektheit zu berücksichtigen. Außerdem sollte es obligatorisch sein, entsprechende Positionen im Unternehmen zu konstituieren, die für das Modell der Datengewinnung und -nutzung zuständig sowie für die Kontrolle der ordnungsgemäßen Umsetzung verantwortlich sind [2].

Vollständigkeit der Daten
Die Grundvoraussetzung für Vollständigkeit ist, dass ein Datensatz alle notwendigen Attribute und Inhalte enthält. Kann die Vollständigkeit der darin inkludierten Daten nicht garantiert werden, entsteht daraus potenziell das Problem von irreführenden Tendenzen, was zu falschen oder diskriminierenden Ergebnissen führt.

Repräsentativität der Daten
Die Repräsentativität zeichnet sich dadurch aus, dass die Daten eine tatsächliche Grundgesamtheit und somit entsprechend die Realität abbilden, die stellvertretend im Sinne der Aufgabenstellung ist. Sind die Daten nicht repräsentativ, hat dies zur Folge, dass daraus ein Bias (Datenverzerrung) resultiert. Ein Bias kann z. B. durch einen Fehler bei der Datenerhebung entstehen.

Nachvollziehbarkeit der Daten
Für die Überprüfung der Datenqualität ist es essenziell, dass nachvollzogen werden kann, aus welchen Quellen die verwendeten Daten stammen. Sind die Quellen nicht transparent, das heißt nicht nachvollziehbar, ist es nicht möglich eine notwendige Validierung der Daten vorzunehmen, was sich auf deren Qualität negativ auswirken kann. Für eine bestmögliche Bewertung und Messung sowohl der Datenqualität als auch der Qualität der Quellen sowie der Ableitung gezielter Verbesserungsmaßnahmen, müssen im Vorfeld entsprechend Vorgaben definiert werden. Hierfür gilt es, die für den Prozess relevanten Kriterien zu bestimmen, etwa Konsistenz oder Einheitlichkeit. Anhand der gewählten Kriterien ist es dann möglich, die erhobenen Daten bezüglich ihrer konsistenten Qualität zu überprüfen.

Aktualität der Daten
Die grundsätzliche Idee beim Maschinellen Lernen oder KI ist die Extraktion von Wissen aus Daten. Aus diesem Grund muss sichergestellt werden, dass die generierten, respektive verwendeten Daten auch die passenden Informationen und Erfahrungen enthalten, um mit den KI-Algorithmen das Problem richtig und vertrauenswürdig zu lösen. Nicht zuletzt aufgrund der Tatsache, dass Menschen sich nicht linear verhalten, können veraltete Daten zu falschen Ergebnissen führen. Aus diesem Grund sollten – in Abhängigkeit von der Anwendung – möglichst aktuelle Daten verwendet werden.

Korrektheit der Daten
Die Daten müssen mit der Realität übereinstimmen und damit für die Anwendung korrekt sein. Die Auswahl der Daten bedingt, dass diese Anforderungen mit einer detaillierten Analyse ermittelt wurden – als Methode kann hier das Mapping gegen Daten, deren Korrektheit bestätigt ist, oder definierte, abgestimmte Plausibilitätsregeln eingesetzt werden. So lässt sich sicherstellen, dass zwischen den – zur Entwicklung oder im Weiteren in der Anwendung – genutzten Daten und der Realität keine Diskrepanz besteht.

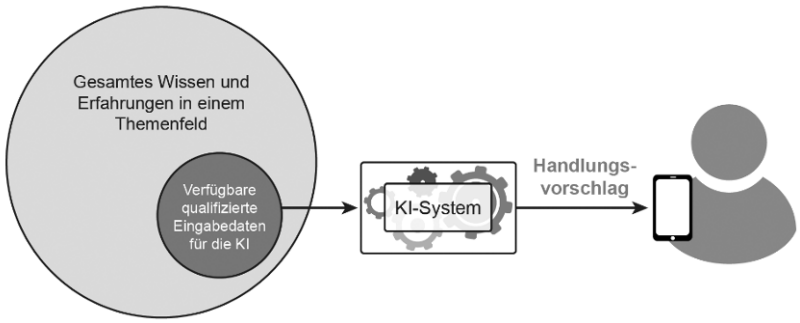
Weitere Aspekte, zur Verbesserung der Datenqualität
Zur Umsetzung einer hohen Datenqualität sind auch die Entwicklung von qualitativ hochwertigen Sensoren sehr hilfreich. Auch Daten aus anderen Quellen müssen sorgfältig und möglichst nachvollziehbar erhoben und vor der Nutzung verifiziert werden. Darüber hinaus helfen hochwertige Datenpools, den Austausch von Daten zu motivieren, eine Interoperabilität von Daten zu schaffen und Open Data-Strategien voranzutreiben, um bessere Ergebnisse zu erreichen.

3 Umgang mit den KI-Ergebnissen

Bei dem Prinzip „Keep the human in the loop" werden die KI-Ergebnisse als Handlungsvorschlag für den Nutzer verstanden (Abb. 3). Dieses fördert die Selbstbestimmtheit des Nutzers und erhöht das Vertrauen in KI-Systeme. Das Ergebnis der KI ist der Handlungsvorschlag für den Nutzer. Bei dieser Vorgehensweise kann der Nutzer das Ergebnis mit seinem individuellen Wissen, seinen Erfahrungen und eigenen Zielen / Werten abgleichen und entscheiden sowie die Verantwortung übernehmen.

Wichtig zu berücksichtigen ist, dass Handlungsvorschläge immer fehlerbehaftet sein können. Bei ChatGPT sind nur 85 der Ergebnisse in bestimmten Bereichen korrekt. Aus diesem Grund muss der Nutzer in der Lage sein, den Handlungsvorschlag über-

Abb. 3 | Keep the human in the loop



prüfen zu können. In Themenfeldern, von denen der Nutzer keine Ahnung hat, ist der schwierig und er sollte das KI-System besser nicht nutzen.

Automatisierte KI-Anwendungen
Bei automatisierten KI-Anwendungen wie z. B. autonomes Fahren, kann das Prinzip „Keep the human in the loop" nicht verwendet werden. Aus diesem Grund muss hier ein hoher Aufwand in Simulation, Test und Validierung gesteckt werden, um qualitativ höherwertige KI-Ergebnisse zu garantieren. Außerdem spielen bei automatisierten KI-Anwendungen Verantwortung, Haftung und Versicherung eine besondere Rolle.

4 Positive und negative Seiten der KI bezüglich IT-Sicherheit

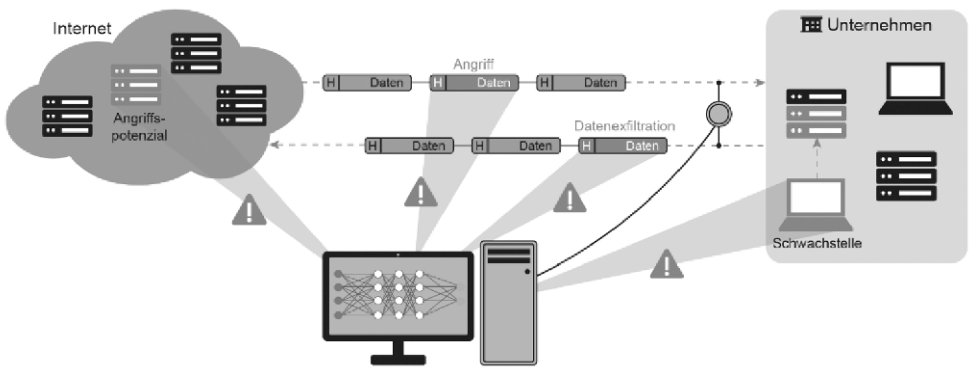
4.1 Positive Nutzung der KI bezüglich der IT-Sicherheit

Die Nutzung von KI für IT-Sicherheit schafft deutliche Mehrwerte für den Schutz von Unternehmen und Organisationen. Im Folgenden werden einige Anwendungsfelder exemplarisch dargestellt.

Erhöhung der Erkennungsrate von Angriffen
Ein erstes wichtiges Themenfeld ist die Erhöhung der Erkennungsrate von Angriffen (Abb. 4).

Es geht z. B. um das Erkennen von Angriffen über das Netzwerk, in den Endgeräten, Servern, IoT-Geräten und Cloudanwendungen. Dazu werden adaptive KI-Modelle benötigt, um auch kontinuierlich neue Angriffsvektoren und Bedrohungen frühzeitig erkennen zu können. Wichtig ist hier aber auch, dass die notwendigen sicherheitsrelevanten Daten aus den Netzwerken und IT-Systemen bekommen, damit die KI daraus nützliche Ergebnisse zur Verbesserung des Schutzniveaus erzielen kann. Weitere Bereiche, bei denen die verbesserte Erkennung eine wichtige Rolle spielt, sind z.B. das Erkennen von Malware, Spam, Fake-News und Deep-Fake.

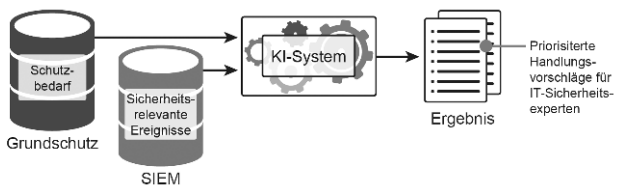
Abb. 4 | Erkennung von Angriffen



Unterstützung / Entlastung von IT-Sicherheitsexperten durch Automatisierung
Ein weiteres relevantes Themenfeld bei dem KI für mehr IT-Sicherheit genutzt werden kann, ist die Unterstützung / Entlastung von IT-Sicherheitsexperten, von denen wir zurzeit viel zu wenige haben.

Priorisierung von IT-Sicherheitsereignissen
Ein erstes Beispiel ist das Analysieren von wichtigen sicherheitsrelevanten Ereignissen mit einer Priorisierung, um dem IT-Sicherheitsexperten zeitaufwendige Analysearbeit abzunehmen. Ein KI-System analysiert zum Beispiel die vielen hunderte oder tausende sicherheitsrelevanten Ereignisse von den IT-Systemen und zeigt dann auf, nach welchen Prioritäten diese vom IT-Sicherheitsexperten abgearbeitet werden müssen, um den höchsten Schutz in der aktuellen Situation für das Unternehmen zu erzielen. In der Abb. 5 ist zu sehen, dass dazu die sicherheitsrelevanten Daten aus einem SIEM-System und die Ergebnisse einer Schutzbedarfsfeststellung verwendet werden können, um eine optimale Priorisierung zu erreichen.

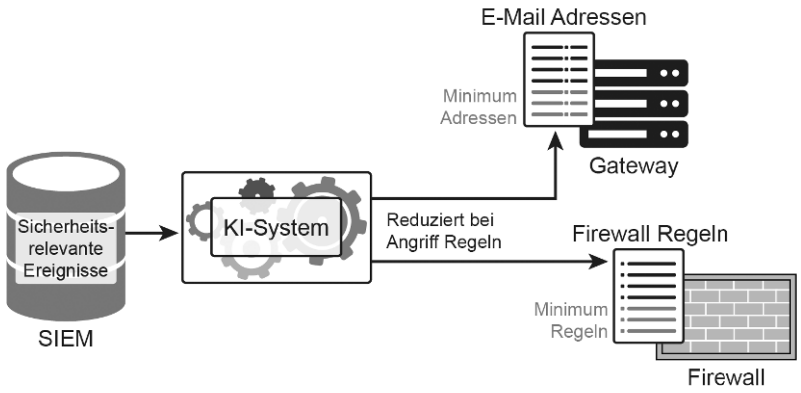
Abb. 5 | Automatisches Erkennen und Priorisieren



(Teil-)Autonomie bei Reaktionen
Beim zweiten Beispiel werden bei der Erkennung eines Angriffes oder einer besonderen Bedrohung sofort Firewall- und E-Mail-Regeln automatisch so reduziert, dass die wichtigen Prozesse für ein Unternehmen aufrechterhalten, bleiben (Siehe Abb. 6 - Minimum Regeln). Dadurch wird die Angriffsfläche für die Angreifer deutlich reduziert, damit Schäden verhindert werden.

Es gibt aber auch noch weitere IT-Sicherheitsbereiche, bei denen KI - auch mit ChatGPT oder ChatGPT-ähnlichen Lösungen - den IT-Sicherheitsexperten helfen kann wie zum

Abb. 6 | Automatische Reaktion bei Angriffen



Beispiel sichere Softwareentwicklung, IT-Forensik und Threat Intelligence.

4.2 Negativ: Die Angreifer werden noch effizienter und erfolgreicher mit ihren Angriffen

Auch die Angreifer nutzen KI, um ihre Angriffe erfolgreicher umsetzen zu können. Im Folgenden werden ein paar prinzipielle Angriffe aufgezeigt:

Angriffsvektoren identifizieren
Schwachstellen / Sicherheitslücken werden mithilfe von KI bei den Opfer IT-Systemen aufgedeckt, um Angriffsmöglichkeiten / -potentiale abzuschätzen und Angriffe effektiv umzusetzen.

IT-Sicherheitsmaßnahmen überwinden
Schwellwerte oder Sequenzen eines Angriffserkennungssystems werden z. B. mithilfe von KI so analysiert, um mit den Ergebnissen Angriffe unentdeckt umzusetzen.

Automatisiertes Angreifen
Es werden Social Bots zur automatischen Generierung von Fake-Accounts und Fake-News zur Manipulation von Menschen verwendet. Mit KI werden künstliche Fotos von Menschen generiert und Profile so konstruieren, dass diese in die gewünschten Filterblasen und Echokammern kommen, um entsprechend und zielgerichtet manipulieren zu können. Auch die eigentlichen Informationen, die manipulieren sollen, können mit KI einfacher und mit qualitativeren Ergebnissen generiert werden.

Social Engineering
LLMs haben die Fähigkeit, den Schreibstil einer bestimmten Person nachzuahmen, wenn sie mit spezifischen Trainingsdaten darauf trainiert wurden. Dies macht Social-Engineering-Betrug, insbesondere Spear-Phishing, einfacher.

Wenn ausreichend persönliche Daten vorhanden sind, kann man mit LLMs Texte im Stil des anvisierten CEOs verfassen. Dies eröffnet neue Möglichkeiten bezüglich CEO-Frauds, denn diese generierten Texte könnten zusammen mit Audio-Imitationen und Deepfake-Videos verwendet werden, um den Angriff auch über Anrufe oder Video-Calls umzusetzen [3].

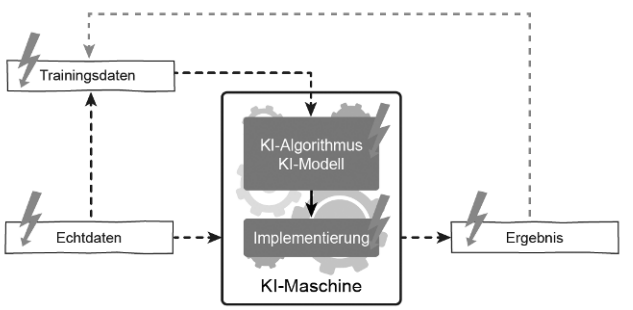
e) Beschleunigtes Entwickeln von Angriffen
LLMs wie ChatGPT können für die Erstellung von Angriffstechnologien wie Malware als nützliches Werkzeug dienen. Das Risiko besteht darin, dass Cyberkriminelle dadurch Angriffe schneller entwickeln können.
Ein weiteres Risiko für die IT-Sicherheit besteht darin, dass ChatGPT bei der Entwicklung von polymorpher Malware eingesetzt werden kann. Polymorphe Malware ist eine Schadsoftware, die ihre Implementierung (Code) verändert, aber die Funktionalität – Schadfunktionen – beibehält. Dies kann durch wiederholte Programmieranfragen mithilfe von ChatGPT erreicht werden. Ändert sich die Malware jedes Mal, da sie polymorph ist, funktioniert die Signatur-Methode nicht mehr. Als Folge davon sinkt die Erkennungsrate der Anti-Malwarelösung.

Vertrauliche Informationen als Input eines zentralen Dienstes
Wenn vertrauliche Informationen in KI-Systeme eingegeben werden, entsteht ein weiteres IT-Sicherheitsrisiko. Es besteht die Möglichkeit, dass diese eingegebenen Daten von KI-Systemen für weiteres Training verwendet werden. Zusätzlich stellt die Eingabe vertraulicher Informationen eine neue Angriffsfläche der eingegebenen Informationen dar.

5 Schutz der KI-Systeme ist notwendig

Elementar ist aber auch der Schutz von KI-Anwendungen in allen Bereichen, wo KI genutzt wird. Angreifer versuchen die Trainingsdaten und Inputdaten sowie Algorithmen und Modelle zu manipulieren, um falsche Ergebnisse von KI-Systemen zu provozieren (Abb. 7). Zum Beispiel im medizinischen Bereich, in der Produktion oder beim autonomen Fahren. Das kann katastrophale Folgen haben.

Abb. 7 | Angriffe auf die Künstliche Intelligenz



Daher muss dafür gesorgt werden, dass die Daten, Algorithmen und Modelle gegen Manipulationen mithilfe von IT-Sicherheitsmaßnahmen geschützt werden.

Schutzziele die beachtet und mit passenden IT-Sicherheitsmaßnahmen umgesetzt werden müssen, sind:

Integrität
Es müssen IT-Sicherheitsmaßnahmen zum Erkennen von Manipulation der Daten umgesetzt werden. Zum Beispiel durch die Berechnung von kryptographischen Prüfsummen, die von der Nutzung überprüft werden.

Vertraulichkeit
Wenn in den Daten wie Trainingsdaten und Echtdaten Geschäftsgeheimisse enthalten sind, muss deren Schutz gewährleistet werden. Zum Beispiel durch die Verschlüsselung der Daten während der Speicherung.

Datenschutz
Wenn personenbezogene Daten enthalten sind, müssen auch diese nach dem Stand der Technik geschützt werden.

Verfügbarkeit
Wenn zunehmend KI-Systeme in der Digitalisierung genutzt werden, wird die Verfügbarkeit immer wichtiger. Dazu müssen die Daten und KI-Systeme entsprechend redundant vorgehalten werden.

Manipulieren von Trainingsdaten (Poisoning Attack)
Das Manipulieren der Trainingsdaten ist ein Angriff auf ein KI-System, um das Modell zu beeinflussen (z. B. Genauigkeit verschlechtern). Insbesondere Anwendungen der IT-Sicherheit sind dagegen anfällig, weil z. B. ein Angreifer Spam erstellt und somit die Trainingsdaten kontrolliert. Spam kann so erstellt werden, dass legitime Inhalte mit Spam assoziiert werden, sodass legitime Mails als Spam klassifiziert werden. Andersherum kann ein Angreifer legitime Mails versenden und bestimmte Inhalte einfügen, die er in einer speziellen Spam-E-Mail nutzen will, sodass diese Inhalte als legitim gelernt werden. Die Spam-E-Mail wird folglich nicht blockiert. Grundsätzliche Frage bei der Wahl von Trainingsdaten bei Spam: Ist der vorliegende Spam echter Spam (das, was gelernt werden soll) oder eine Poisoning Attack, die wie Spam aussieht, um das Modell zu manipulieren? Online Learning Systems lernen durchgehend und sind dadurch anpassbar an sich ständig verändernde Angriffe, dadurch aber auch anfällig für Poisoning (bzw. causative) Attacks.

Allgemein fügt bei einer Poisoning Attack ein Angreifer bösartige Samples in die Trainingsdaten ein, um die Entscheidungen der betroffenen KI-Systeme zu beeinflussen. Die Voraussetzung für diesen Angriff ist ein Zugriff auf die Trainingsdaten entweder mittelbar (z. B. Feedback-Loop) oder unmittelbar. Ziel des Angreifers ist es, das KI-Modell so zu verändern, dass es zu seinen Gunsten falsche Entscheidungen trifft.

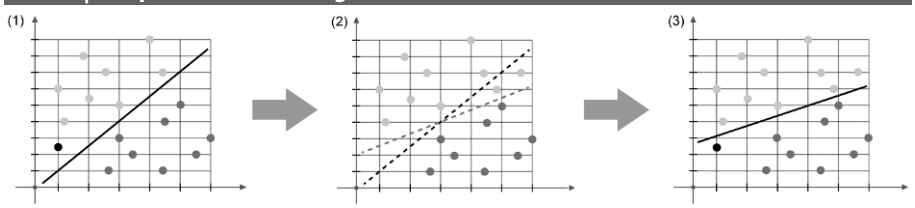
Beispiel einer Poisoning Attack (Abb. 8): In diesem Beispiel werden bei der Support-Vector-Machine die klassifizierten Eingangsdaten so modifiziert, dass die Hyperebene zur Trennung der klassifizierten Objekte so verändert wird, dass dadurch gezielt unerkannte Angriffe möglich sind.

Manipulieren der Echtdaten (Evasion Attack)
Erstellung einer Eingabe, eines sogenannten Adversarial Examples, dass eine falsche oder bestimmte Vorhersage/Klassifizierung verursachen soll. Oft dient eine natürliche Eingabe als Ausgangspunkt und wird gezielt modifiziert: Einfügen von speziellen Störungen/Rauschen (Perturbation/Distortion), die für einen Menschen nicht sichtbar sind. Adversarial Examples können aber auch in der Praxis vorkommen, zum Beispiel ein teilweise bemaltes oder beklebtes Verkehrsschild (es kommt auf die Sichtweise an, ob solche Fälle natürlichen oder bösartigen Ursprungs sind; da beklebte Schilder oft vorkommen, sollten sie eigentlich Teil der Trainingsdaten sein). Ein anderes Beispiel ist eine Brille, um sich gegenüber einer Gesichtserkennung zu verstecken oder sich als eine bestimmte Person auszugeben (Backdoor Attack: ein spezieller Trigger wird mit einer bestimmten Klasse assoziiert, ansonsten soll das Modell nicht weiter verändert werden). In der Regel benötigt eine Evasion Attack Zugriff auf Eingabe-Ausgabe-Paare des Ziel-Modells (ein Prediction-Interface / Orakel), um zu testen welche Angriffsvektoren funktionieren und mit welcher Konfidenz (wenn das Modell Konfidenzwert zurückgibt). Oft geschieht dies in Kombination mit einer Model Extraction Attack, bei der ein ähnliches Modell nachgebaut wird und die gesammelten Eingabe-Ausgabe-Paare erlauben, die Ähnlichkeit des lokalen Modells zu dem Ziel-Modell einzuschätzen. Die Transferability Property beschreibt die Eigenschaft, dass ein Adversarial Example, das für ein Modell funktioniert, mit hoher Wahrscheinlichkeit auch für ein anderes Modell funktionieren wird, wenn beide Modelle für die gleiche Aufgabe trainiert wurden.

Generell erzeugt bei einer Evasion Attack ein Angreifer speziell gestaltete Eingaben, um eine falsche Entscheidung bei einem KI-Modell zu verursachen, um zum Beispiel einer Detektion zu entgehen. Diese Eingaben können so gestaltet sein, dass sie von Menschen als normal wahrgenommen werden, aber von KI-Algorithmen falsch klassifiziert werden.

6 Der Einsatz von KI basiert auf Vertrauenswürdigkeit

Abb. 8 | Beispiel einer Poisoning Attack



(1) Normale Klassifizierung eines neuen Inputs. (neuer schwarzer Punkt gehört zur hellgrauen Klasse)
(2) Beispiel: Manipulation von Trainingsdaten; Falsch klassifizierte Daten werden in den Trainingsprozess als Angriff eingeschleust (zwei weitere hellgraue Punkte). Dadurch wird die Gerade des Modells zur Klassifizierung manipuliert (Gerade wird flacher).
(3) Damit kann ein Angreifer für falsche Klassifizierungen sorgen (jetzt gehört der neue schwarze Punkt zur dunkegrauen Klasse).

Da nicht nur aus dem Einsatz von KI neue Schwachstellen resultieren, sondern auch mittels KI neue Angriffsmöglichkeiten auf Staat und Unternehmen realisierbar sind, erscheint es als eine logische Schlussfolgerung kriminelle Attacken auf gleicher Ebene abzuwehren – also mithilfe von KI den Schutz von IT-Systemen und IT-Infrastrukturen zu gewährleisten. Obwohl KI im Kontext der IT-Sicherheit unweigerlich als probates Mittel zur Abwehr scheint, ist auch hier – wie bei

jedem Einsatz der KI – die Frage zu diskutieren, in welchem Rahmen dies angemessen ist [4].

Warum dieser Diskurs nötig ist und welche Forschungsfragen in diesem Kontext zu bearbeiten sind, wird nachfolgend exemplarisch anhand von zwei Szenarien aus dem Bereich des Schutzes der IT dokumentiert.

Privatheit vs. Allgemeinwohl: Problem Schutz von Daten
Die Berücksichtigung der Annahme, dass mittels KI eine Optimierung der IT-Schutzmaßnahmen realisierbar ist, bedeutet, dass ein Angriff schneller und präziser detektiert werden kann, je mehr Daten mit sicherheitsrelevanten Informationen zur Verfügung stehen. Daraus lässt sich im Weiteren auch ableiten, dass für eine bestmögliche Kennung in bestimmten Fällen die Einbeziehung personenbezogener Daten von Mitarbeitern unverzichtbar ist, da viele Angriffe mithilfe von Social Engineering und Malware über die Endgeräte der Mitarbeiter indirekt durchgeführt werden. Eine Notwendigkeit für deren Hinzunahme könnte sich zum Beispiel für den Fall ergeben, wenn ein Energieversorger angegriffen wird, da dies eine präzise Ursachenforschung für die schnellstmögliche Reaktion zur Gewährleistung der Stromversorgung verlangt. Der Bedarf ist theoretisch nachvollziehbar – andererseits werden dadurch die Individualrechte außer Kraft gesetzt, da die Nutzung von personenbezogenen Daten durch die DSGVO exakt limitiert ist. Im konkreten Fall würde dies bedeuten, dass durch eine Analyse der Daten parallel das Verhalten der Mitarbeiter ausgewertet und somit deren Privatsphäre verletzt wird.

Daraus resultiert folgendes Dilemma: Die Ethik fordert, dass keine grundlegenden Rechte wie die Privatheit zugunsten eines höheren Ziels völlig aufgegeben werden dürfen. Dagegen steht gemäß dem Utilitarismus die Prämisse des Strebens nach dem Gesamtnutzen für die Gesellschaft. Aus dieser resultiert dann zwangsläufig die Fragestellung, wann es angeraten oder sogar unabdingbar ist die Rechte des Individuums zugunsten des Wohles für die Gemeinschaft aufzuheben. Eine mögliche Annäherung zur Auflösung dieses Dilemmas könnte sein, hierfür Grenzen zu definieren, indem Kriterien dafür festgelegt werden, wann das Individualrecht nachrangig zu behandeln ist.

Strike Back: Problem Unvollständigkeit der Daten
Dem Konstrukt des ‚Strike Back' liegt die Hypothese zugrunde, dass ein Angriff durch einen Gegenangriff beendet werden kann. Unter Einsatz von KI wäre es theoretisch möglich, eine vollkommen neutrale Einschätzung zu ermitteln, was unternommen werden müsste, um den Angreifer zum Aufgeben zu motivieren. Die ethische Frage in diesem Kontext dreht sich um den Begriff der Neutralität und ob es überhaupt möglich ist, die Vollständigkeit der Daten zu erreichen, um eine ausgewogene und verantwortliche Entscheidung durch KI-Systeme errechnen zu lassen. Davon ausgehend, dass ein Strike Back automatisiert erfolgt, könnte möglicherweise ein Schaden auf einer höheren Ebene angerichtet werden, der moralisch nicht vertretbar wäre – wie etwa ein Gegenschlag, der gleichzeitig auch die Stromversorgung eines Krankenhauses lahmlegen würde. Fragestellungen wie diese zeigen den eklatanten Forschungsbedarf dahingehend auf, inwieweit es leistbar ist vertrauenswürdige KI-Systeme zu entwickeln, die den Schutz der Zivilgesellschaft gewährleisten können.

Die noch bestehende Unklarheit, ob KI-Systeme die perfekte Lösung sind, sollte zu dem Schluss führen, dass es insbesondere aufgrund der Komplexität sowie der schnellen Weiterentwicklung angemessen ist, grundsätzlich bei wichtigen Anwendungen den Menschen als kontrollierenden Faktor gemäß dem Grundsatz „Keep the human in the loop" einzubinden.

7 Fazit und Ausblick

Künstliche Intelligenz ist eine essenzielle Technologie im Bereich der IT-Sicherheit.

Erkennen von Angriffen, Bedrohungen, Schwachstellen, Fake-News sowie die Unterstützung / Entlastung von IT-Sicherheitsexperten aber auch sichere Software-Entwicklung und weitere Felder spielen eine relevante Rolle.

Klar ist, dass Hacker Daten, Algorithmen/Modelle angreifen, um Ergebnisse zu manipulieren. Aus diesem Grund ist der Schutz von KI-Systemen von hoher Bedeutung.

Die Angreifer nutzen KI sehr professionell und erfolgreich für ihre Angriffe. Jetzt sollten die Verteidiger KI deutlich intensiver und flächendeckend nutzen, um sich besser verteidigen zu können.

Referenzen

[1] N. Pohlmann: „Cyber-Sicherheit – Das Lehrbuch für Konzepte, Mechanismen, Architekturen und Eigenschaften von Cyber-Sicherheitssystemen in der Digitalisierung", Springer-Vieweg Verlag, Wiesbaden 2022

[2] U. Coester, N. Pohlmann: „Diskriminierung und weniger Selbstbestimmung? Die Schattenseiten der Algorithmen", tec4u, 12/17

[3] P. Farwick, N. Pohlmann: „Chancen und Risiken von ChatGPT – Vom angemessenen Umgang mit künstlicher Sprachintelligenz", IT-Sicherheit – Mittelstandsmagazin für Informationssicherheit und Datenschutz, DATAKONTEXT-Fachverlag, 4/2023

[4] U. Coester, N. Pohlmann: „Ethik und künstliche Intelligenz - Wer macht die Spielregeln für die KI?", IT & Production – Zeitschrift für erfolgreiche Produktion, TeDo Verlag, 2019