Künstliche Intelligenz in Bahnanwendungen – neue Angriffsvektoren und Schutzmechanismen

Künstliche Intelligenz nimmt Einzug in immer mehr Bereiche des Lebens, darunter auch sicherheitsrelevante Bahnanwendungen. In diesem Beitrag wird darauf eingegangen, wie ein Einsatz von KI zur Erhöhung der Sicherheit gegen unberechtigte Zugriffe Dritter beitragen kann. Weiterhin wird diskutiert, welche potenziellen Sicherheitsprobleme durch den Einsatz von KI entstehen und wie diese auch wieder verteidigt werden können. Das Ziel ist es, einen fundierten Überblick über die wichtigsten Schutzmechanismen und Angriffsvektoren zu erhalten, die der Einsatz von KI in Bahnanwendungen bringen kann.

1. Einleitung

Seit einiger Zeit beherrscht künstliche Intelligenz (KI) die Medienlandschaft zunehmend: Zunächst als mögliche Schlüsseltechnologie erkannt, schreitet die Forschung zu und damit die Fähigkeiten von künstlicher Intelligenz immer weiter voran. Auch in Bahnanwendungen werden die Potenziale der künstlichen Intelligenz in nahezu allen Bereichen erkannt. Der Einsatz von KI kann dabei Vorteile für alle Akteure bringen - aber auch die Gefahr, dass neue Sicherheitslücken entstehen.

Eine Notwendigkeit zum Einsatz von KI kann aus Sicht der IT-Sicherheit entstehen: In diesem Feld existieren traditionell Spannungen zwischen Angreifer und Verteidiger. Nehmen sich die Angreifer nun künstliche Intelligenz zur Hilfe, um Angriffe zu verfeinern, kann schnell ein Ungleichgewicht entstehen, welches unbedingt zu verhindern ist. Diese Übermacht der Angreifer bringt den klassischen Sicherheitszyklus aus dem Gleichgewicht. Eine mögliche Antwort darauf ist der Einsatz von KI auf Seite der Verteidiger.

Insgesamt lässt sich der Bereich KI-Sicherheit also durch zwei Kernfragen charakterisieren: Welche neuen Angriffsvektoren entstehen durch einen Einsatz von KI? Wie kann KI auch für klassische IT/OT-Sicherheit eingesetzt werden?

Die Agentur der Europäischen Union für Cybersicherheit (ENISA) schreibt, dass der Großteil der bisher observierten Angriffe auf Bahnsysteme auf deren IT-Systeme abzielt. Wenn OT-Systeme betroffen waren, dann bisher nur dadurch, dass zuvor sicherheitsrelevante IT-Systeme angegriffen und in der Folge ausgefallen sind [1]. Deshalb ist im Rahmen dieses Beitrags eine Betrachtung der OT- und vor allem auch der IT-Systeme notwendig.

2. Grundlagen

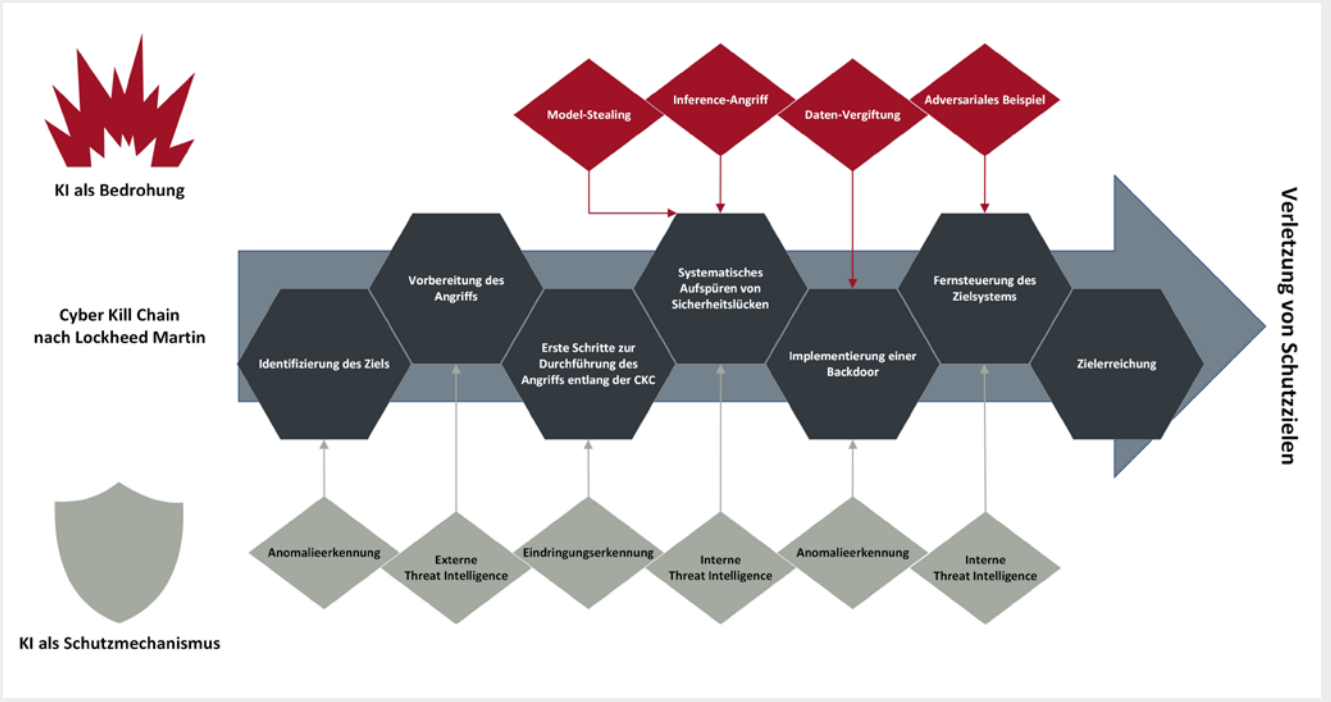
Künstliche Intelligenz ist ein Bereich der Informatik, in dem versucht wird, menschliche Intelligenz in verschiedenen Formen zu imitieren. Maschinelles Lernen ist ein Teilbereich der KI, in dem Modelle entwickelt werden, die aus Daten Muster ableiten können, die vorher nicht explizit programmiert wurden. Dieser Punkt markiert auch einen wichtigen Unterschied zwischen maschinellem Lernen und einem klassischen Algorithmus, der lediglich auf eine Datenbank zur Vorhersage zurückgreifen kann.

Wenn maschinelles Lernen zum Einsatz kommen soll, muss dieses in der Trainingsphase lernen, in Trainingsdaten Muster zu erkennen. Dazu werden vor allem zwei Ansätze verfolgt: Beim überwachten Lernen wird ein Modell auf Daten trainiert, die zuvor mit einer gewünschten Vorhersage markiert wurden. Das Ziel des Modells ist es dann, eine Zuordnung zwischen den Eingabedaten und den Ausgabebezeichnungen zu erlernen, sodass es später die richtige Ausgabebezeichnung für unbekannte Eingabedaten vorhersagen kann. Beim unüberwachten Lernen hingegen existieren die Markierungen der gewünschten Vorhersagen nicht. In einem solchen Fall soll das Modell dann lernen, Muster oder Strukturen in den Eingabedaten zu entdecken.

Die *Cyber Kill Chain* (siehe Bild 1) wurde von Lockheed Martin entwickelt, um den

Jan Malte Hilgefort, M. Sc.
Cyber Security Engineer bei der ESE Engineering und Software-Entwicklung GmbH
jan.malte.hilgefort@ese.de

Prof. Dr.-Ing. habil. Lars Schnieder
Neben seiner Tätigkeit als Geschäftsführer der ESE Engineering und Software-Entwicklung GmbH lehrt er als Privatdozent an der RWTH Aachen und ist an der Technischen Universität Braunschweig zum Honorarprofessor bestellt
Lars.Schnieder@ese.de



1: Risiken (oben) und Chancen (unten) durch KI anhand der Cyber Kill Chain nach Lockheed Martin Quelle: Eigene Darstellung

allgemeinen Ablauf von Cyberangriffen in sieben aufeinander folgenden Phasen besser klassifizieren zu können [2].

3. Risiken: Manipulation von KI durch unberechtigte Dritte

Ein Angriff im Sinne der Cybersicherheit ist jeglicher Versuch, die Vertraulichkeit, Integrität oder Verfügbarkeit von Informationen zu beeinträchtigen. Das Bundesamt für Sicherheit in der Informationstechnik (BSI) benennt die vier wichtigsten KI-spezifischen Angriffe in [3], die nun kurz mit jeweiligen Mitigationsmöglichkeiten vorgestellt werden. In Bild 2 ist der Unterschied zwischen herkömmlichen Angriffen und solchen auf KI dargestellt.

3.1. Adversariale Beispiele

Um bei der Entwicklung die Fehler zu beheben, die in der realen Welt wahrscheinlich am häufigsten auftreten, wird zunächst für gewöhnlich von einem durchschnittlichen Use-Case ausgegangen. Mit dem Einsatz künstlicher Intelligenz stellt sich die Frage, ob diese Vorgehensweise Grenzfälle unbeachtet lässt, in denen die KI anders agiert, als intuitiv anzunehmen wäre. Ein solcher Grenzfall wird adversariales Beispiel genannt und setzt in der sechsten Phase der *Cyber Kill Chain* an. Ein adversariales Beispiel ermöglicht es einem Angreifer, ein Modell so zu beeinflussen, dass potenziell arbiträre Ausgaben zur Fernsteuerung des Systems erzielt werden können.

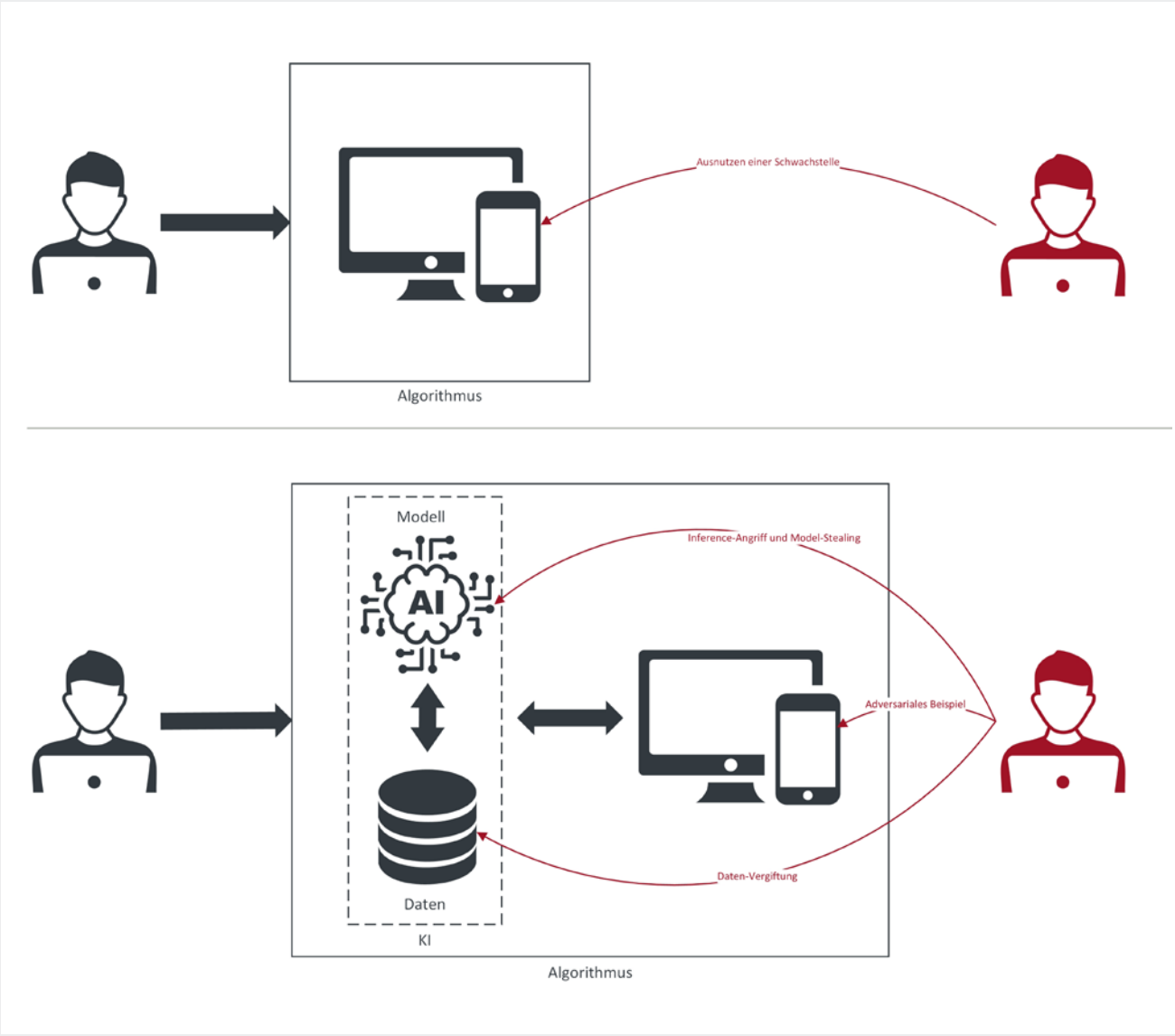
Adversariale Beispiele sind in der Realität auch in Black-Box-Systemen zu finden, indem durch gezielte Eingaben und den daraus resultierenden Ausgaben die Grenzbereiche der Vorhersage der KI gesucht werden. Bei bildbasierten Klassifikatoren kann ein adversariales Beispiel erzeugt werden, indem die einzelnen Pixel des Eingabebildes so lange verändert werden, bis die gewünschte Ausgabe erzielt wurde. Gibt die KI neben der Klassifizierung auch Konfidenzwerte aus, kann ein gieriger AI-gorithmus zunächst bestimmen, welcher Pixel die Ausgabe am meisten in Richtung der gewünschten Ausgabe verändert. Dies kann iterativ so lange erfolgen, bis sich die Ausgabe wie gewünscht verändert. Wird pro Pixel dann noch eine maximale Veränderung festgelegt, kann ein adversariales Beispiel oft erzeugt werden, ohne dass es für einen Menschen als bösartig zu erkennen ist [5].

In einem Beitrag von Huang et al. werden zwei Methoden vorgestellt, möglichst effektiv adversariale Beispiele zu erzeugen, die autonom fahrende Straßenbahnen täuschen sollen. Die Autoren können zeigen, dass sich in unter 4 Minuten ein adversariales Beispiel erzeugen lässt, welches für den Menschen nicht von einem normalen Bild zu unterscheiden ist. Das erzeugte Bild sorgt aber dafür, dass das Computer-Vision-System der Straßenbahn keine Klassifikation von Objekten im Straßenverkehr mehr vornehmen kann [6].

Verteidigt werden kann eine solche Schwachstelle zum Beispiel, indem *Differential Privacy* angewandt wird. Dieses Konzept beschreibt das bewusste Versehen von Trainingsdaten mit Rauschen, um den Einfluss des einzelnen Datenpunkts auf das Training zu reduzieren. Dadurch können die Testdaten zum einen anonymisiert werden, zum anderen sorgt das Rauschen in den Testdaten aber auch für eine erhöhte Robustheit gegen adversariale Beispiele.

3.2. Daten-Vergiftung

Bei Vergiftungsangriffen handelt es sich um eine gezielte Manipulation der Trainingsdaten durch die Entwicklung und Einschleusung von bösartigen Datenpunkten, um den Lernprozess der künstlichen Intelligenz zu kompromittieren. Damit zielt ein solcher Angriff auf die Verletzung der Integrität der KI ab. Daten-Vergiftung wurde in die fünfte Phase der *Cyber Kill Chain* eingeordnet. Die Manipulation der



2: Unterschied zwischen einem klassischen Angriff und einem Angriff auf ein KI-gestütztes System. Angepasst aus [4]

Trainingsdaten kann von einem Angreifer so vorgenommen werden, dass sie der Implementierung einer Hintertür gleicht, die später ausgenutzt werden kann.

Auch wenn Trainingsdatensätze für gewöhnlich nicht der Öffentlichkeit zugänglich sind, wird dieser Angriff in der Realität häufig verwendet [7]. Dies ist möglich, weil Trainingsdaten oft zum Beispiel über Datensammelpunkte wie Honeypots online gesammelt werden und Angreifer mit diesem Wissen gezielt in diese einspielen können.

Eine KI, die auf vergifteten Daten trainiert wurde, ist nicht einfach zu erkennen. In einem bekannten Beispiel ist eine für das autonome Fahren ausgelegte KI durch eine Daten-Vergiftung so trainiert worden, dass ein bestimmt gefärbter Aufkleber eine falsche Klassifizierung erzeugt. Wird dieser Aufkleber auf einem Stoppschild in der realen Welt angebracht, klassifiziert die KI dieses als Geschwindigkeitsbegrenzung [8]. Ein ähnlicher Angriff beispielsweise auf ein zugseitig verbautes Computer-Vision-System, das GoA3+ ermöglichen soll, ist potenziell lebensbedrohlich.

Vergiftungsangriffe können zum Beispiel verteidigt werden, indem die Trainingsdaten einem professionellen Datenmanagement unterliegen [4].

3.3. Inference-Angriff

Mit Inference-Attacks wird versucht, Informationen über die Trainingsdaten zu erhalten. Ein Inference-Angriff zielt auf eine Verletzung der Vertraulichkeit des Systems und setzt in der vierten Phase der *Cyber Kill Chain* an. Die aufgedeckten Trainingsdaten können vom Angreifer dazu genutzt werden, mehr über das Verhalten des KI-Modells zu erfahren und damit weitere potenzielle Sicherheitslücken zu finden.

Zugriff auf die Trainingsdaten ermöglicht eine genaue Analyse. Vor dem Hintergrund der DSGVO[1)] können hier zum Beispiel Probleme entstehen, wenn Systeme mit persönlichen Daten trainiert werden. Bereits 2015 konnten Frederikson et al. zeigen, dass Gesichter aus einem Trainingsdatensatz wiederhergestellt werden können [9].

1) Datenschutz-Grundverordnung

Werden Trainingsdaten allerdings homomorph verschlüsselt, können die für die Lernphase wichtigen Operationen auf ihnen ausgeführt werden, ein Inference-Angriff kann entsprechend aber nur verschlüsselte Daten offenlegen [10]. Auch die Verwendung von *Differential Privacy* kann zur Anonymisierung der Trainingsdaten verwendet werden.

3.4. Model-Stealing

Model-Stealing-Angriffe beschreiben eine Klasse von Angriffen, die mit Black-Box-Zugriff und ohne Wissen über Parameter oder Trainingsdaten die Funktionalität einer künstlichen Intelligenz kopieren. Manche dieser Angriffe können als Inference-Attack durchgeführt werden und zielen ebenfalls auf die Vertraulichkeit des Systems. Aufgrund der inhaltlichen Nähe zu Inference-Angriffen ist auch Model-Stealing in der vierten Phase der *Cyber Kill Chain* angesetzt. Ein Angreifer kann mit einer vollständigen Rekonstruktion eines KI-Modells noch tiefere Einblicke in das Verhalten gewinnen und damit ebenfalls potenzielle Sicherheitslücken finden.

Neben der reinen Klassifikation geben viele KIs zum Beispiel auch Konfidenzintervalle aus, mithilfe derer ein anderes Netzwerk trainiert werden kann. Durch gezielte Anfragen und die dazugehörigen Rückgaben kann ein Angreifer das zugrunde liegende Modell lokal rekonstruieren.

Diese Angriffe können reduziert werden, indem weniger Daten zurückgegeben werden. Eine maximale Anfragerate oder eine Erkennung wiederholt ähnlicher Eingaben kann ebenfalls ein effektiver Schutz gegen adversariale Beispiele sein. Ganz eliminiert werden kann das Restrisiko nicht - hier ist eine sorgfältige Abwägung der Interessen notwendig [11].

4. Chancen der KI im Schutz gegen unberechtigte Zugriffe Dritter

4.1. Eindringungserkennung (Intrusion Detection)

Ein Sicherheitskonzept, mit dem viele Menschen auch im Alltag bewusst Berührungspunkte haben, ist die Eindringungserkennung. Die überwiegende Mehrheit der Deutschen verwendet Virenscanner auf ihren Systemen, die häufig als sogenannte Intrusion-Detection-Systeme (IDS) ganzheitlich gegen Angriffe von unberechtigten Dritten schützen sollen. Die Eindringungserkennung wird in die dritte Phase der *Cyber Kill Chain* eingeordnet. In diesem Schritt wird vom Angreifer versucht, erste Schadprogramme in das System zu schleusen, zum Beispiel über bösartige E-Mails oder USB-Sticks. KI-gestützte Eindringungserkennung kann hier effektiv verhindern, dass die Schadprogramme tatsächlich in das System gelangen.

Zwei große Nachteile von nicht-KI-basierter Eindringungserkennung, die vor allem mit manuell erstellten Regeln arbeitet, sind die inhärente Nachzeitigkeit, die dadurch entsteht, dass der Schadcode vor der Erstellung der konkreten Maßnahmen dagegen erst einmal bekannt sein muss. Außerdem ist die händische Analyse sehr zeitaufwendig und skaliert somit schlecht. Wird eine KI dagegen zum Beispiel auf die Erkennung bösartiger Signaturen trainiert, kann diese potenziell zugrunde liegende Konzepte erlernen, die dann eine bessere Erkennung von obfuskierten (verschleierten) Daten ermöglichen [12].

4.2. Anomalieerkennung (Anomaly Detection)

Ein Ansatz zur Eindringungserkennung ist die Anomalieerkennung. Hierzu wird die zentrale Fragestellung umgekehrt: Es wird nicht versucht, die Angriffe zu detektieren und charakterisieren, sondern gutartige Datenflüsse möglichst gut zu modellieren, um Anomalien in diesen als wahrscheinliche Angriffe zu erkennen. Eine grundlegende semantische Lücke ist dabei zu sehen: Nicht jede Anomalie muss auch gleich ein Angriff sein. In der *Cyber Kill Chain* kann Anomalieerkennung vor allem an zwei Stellen bei der Abwehr eines Angriffs helfen. In der ersten Phase sucht ein Angreifer nach potenziellen Zielen für einen Angriff. Eine Anomalieerkennung kann dabei zum Beispiel eine ungewöhnlich hohe Anfragerate im Netzwerk feststellen und damit einen Indikator für einen eventuell bevorstehenden Angriff darstellen. Die fünfte Phase ist dadurch geprägt, dass der Angreifer einen Weg finden muss, sich unbemerkt Zugriff auf das System von außen zu schaffen. Eine Anomalieerkennung kann diese Hintertür, Programme entlang der Implementation oder Kommunikationen nach außen erkennen und melden.

Da im Regelfall der Anteil gutartiger Datenpunkte in einem Datensatz deutlich größer als der bösartiger ist, kann unüberwachtes Lernen zum Einsatz kommen, um die nicht klassifizierten Trainingsdaten zu

modellieren. Hier stellt sich die naheliegende Frage, ob ein Training auf einem solchen Datensatz nicht auch die bösartigen Daten als gutartig modellieren würde. Dies ist nicht der Fall, wenn der Anteil bösartiger Daten sehr gering und die Daten selbst klar von gutartigen unterscheidbar sind. In diesem Fall helfen die Daten sogar dabei, ein klassisches Problem von KIs, das Overfitting, zu minimieren. Wang et al. haben eine Anomalieerkennung für den Einsatz im Bahnbereich entwickelt. Sie zeigen, dass sich die Technologie gut dafür eignet, zum Beispiel die Klimaanlagensteuerung, das Traktionssystem oder die Türsteuerungen einer Metro zu überwachen [13].

4.3. Threat Intelligence

Die Threat Intelligence ist ein Sammelbegriff für alle Informationen, die über potenzielle oder aktuelle Bedrohungen bekannt sind. Das Sammeln solcher Informationen und die Einordnung dieser in den anwendungsbezogenen Kontext ist eine Aufgabe von Cybersecurity-Experten, die bei der steigenden Zahl der Bedrohungen immer wichtiger und umfassender wird. Zur Unterstützung bei der Entscheidungsfindung gerade bei taktischer und operativer Threat Intelligence bieten sich künstliche Intelligenzen an, die Analyseschritte übernehmen.

Bei der taktischen Threat Intelligence, die zum Beispiel mögliche Angriffspunkte ausfindig macht, kann künstliche Intelligenz zur Texterfassung zum Einsatz kommen. Die operative Threat Intelligence dagegen beschäftigt sich mit der konkreten Ausgestaltung der Umsetzung der taktischen Ziele. Künstliche Intelligenz kann hier als Unterstützungshilfe ebenso wie als konkrete Maßnahme verwendet werden – zum Beispiel zur Anomalieerkennung [14].

Im OT-Kontext können zum Beispiel Drohnen zur Threat Intelligence eingesetzt werden, die KI-gestützt autonom Informationen über potenzielle Bedrohungen sammeln, die zum Beispiel aus Änderungen an der Bahninfrastruktur oder der bloßen Erkennung Unberechtigter in Sperrbereichen antizipiert werden können.

Externe Threat Intelligence kann in der zweiten Phase der *Cyber Kill Chain*, in der ein Angreifer bestimmt, welche Werkzeuge er für den Angriff verwenden wird, die Menge der möglichen Werkzeuge mitsamt ihren jeweiligen Stärken und Schwächen identifizieren, was die Chance auf Prävention eines Angriffs erhöht. Interne Threat Intelligence zielt dagegen auf die Erkennung von Gefahren durch interne Quellen. Die vierte und sechste Phase kann davon potenziell profitieren, indem hier Sicherheitslücken im eigenen System durch interne Threat Intelligence aufgedeckt und daraufhin geschlossen werden oder die Infrastruktur zur Fernsteuerung des Systems gefunden und dann beseitigt wird.

5. Zusammenfassung und Ausblick

Der vermehrte Einsatz von KI auch für sicherheitsrelevante Bahnanwendungen ist zwar nötig, stellt aber auch eine disruptive Innovation dar. Wie dieser Beitrag gezeigt hat, entstehen durch den Einsatz neue Angriffsvektoren, für die neben einer Verteidigung zunächst auch Bewusstsein geschaffen werden muss. Auf der anderen Seite können KI-basierte Methoden auch bei der Erkennung und Verteidigung von Angriffen und Schwachstellen helfen. Die Vielseitigkeit von KI-basierten Angriffen und Verteidigungen konnte zum Beispiel daran veranschaulicht werden, dass an unterschiedlichsten Stellen entlang der *Cyber Kill Chain* angesetzt wird.

Die Einordnung in den KI-Lebenszyklus hat gezeigt, dass Manipulation sowohl vor, als auch nach der Inbetriebnahme möglich ist. Damit deckt sich die Einschätzung der befragten Deutschen, die sich zu 81 % eine unabhängige Prüfung der Sicherheit von KI-gestützten Produkten und Anwendungen vor der Markteinführung wünschen. Auch nach der Inbetriebnahme sind 79% der Befragten der Meinung, dass unabhängige Prüfungen weiterhin erforderlich sind [15].

Eine aussagekräftige Prüfung erfordert dabei aber immer einheitliche Standards. Der Bedarf an internationalen Normen und Standards für KI wurde dazu in der „Normungsroadmap Künstliche Intelligenz“ des Deutschen Instituts für Normung in der mittlerweile zweiten Fassung ermittelt, in der die Sicherheit und Erklärbarkeit von KI ein zentraler Bestandteil ist [16]. •

Literatur

[1] Agentur der Europäischen Union für Cybersicherheit;„ENISA Threat Landscape: Transport Sector,“ 2023.
[2] E. Hutchins, M. Cloppert und R. Amin, Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains, 2011.
[3] Bundesamt für Sicherheit in der Informationstechnik, Secure, robust and transparent application of AI, 2021.
[4] R. Hamon, H. Junklewitz und J. Sanchez Martin, Robustness and Explainability of Artificial Intelligence, Publications Office of the European Union, 2020.
[5] Bundesamt für Sicherheit in der Informationstechnik, Security of AI-Systems: Fundamentals- Provision or use of external data or trained models, 2022.
[6] S. Huang, X. Liu, X. Yang, Z. Zhang und L. Yang,„Two Improved Methods of Generating Adversarial Examples against Faster R-CNNs for Tram Environment Perception Systems,“ Complexity, 2020.
[7] Z. Tian, L. Cui, J. Liang und S. Yu,„A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning,“ ACM Computing Surveys, 2022.
[8] T. Gu, B. Dolan-Gavitt und S. Garg,„BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,“ arXiv:1708.06733, 2019.
[9] M. Fredrikson, S. Jha und T. Ristenpart,„Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures,“ Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, p. 1322–1333, 2015.
[10] D. Syed, S. S. Refaat und O. Bouhali,„Privacy Preservation of Data-Driven Models in Smart Grids Using Homomorphic Encryption,“ Information, 2020.
[11] F. Tramér, F. Zhang, A. Juels, M. Reiter und T. Ristenpart,„Stealing Machine Learning Models via Prediction APIs,“ Proceedings of the 25th USENIX Security Symposium, 2016.
[12] G. Kumar, K. Kumar und M. Sachdeva,„The use of artificial intelligence based techniques for intrusion detection: a review,“ Artificial Intelligence Review, 2010.
[13] Y. Wang, X. Du, Z. Lu, Q. Duan und J. Wu,„Improved LSTM-Based Time-Series Anomaly Detection in Rail Transit Operation Environments,“ IEEE Transactions on Industrial Informatics, pp. 9027-9036, 20222.
[14] R. Trifonov, O. Nakov und V. Mladenov,„Artificial Intelligence in Cyber Threats Intelligence,“ 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), pp. 1-4, 2018.
[15] TÜV-Verband e.V.,„Sicherheit und Künstliche Intelligenz - Erwartungen, Hoffnungen, Risiken,“ Repräsentative Befragung der Bevölkerung in Deutschland im Auftrag des TÜV-Verbands, 2021.
[16] Deutsches Institut für Normung, Deutsche Normungsroadmap Künstliche Intelligenz (Ausgabe 2), 2022.

Summary

Artificial intelligence in railway applications - new attack vectors and protection mechanisms

Artificial intelligence is finding its way into more and more areas of life, including security-relevant railway applications. This article looks at how the use of AI can help to increase security against unauthorised access by third parties. It also discusses the potential security problems arising from the use of AI and how these can be defended against. The aim is to provide a well-founded overview of the most important protection mechanisms and attack vectors that the use of AI in railway applications can bring.