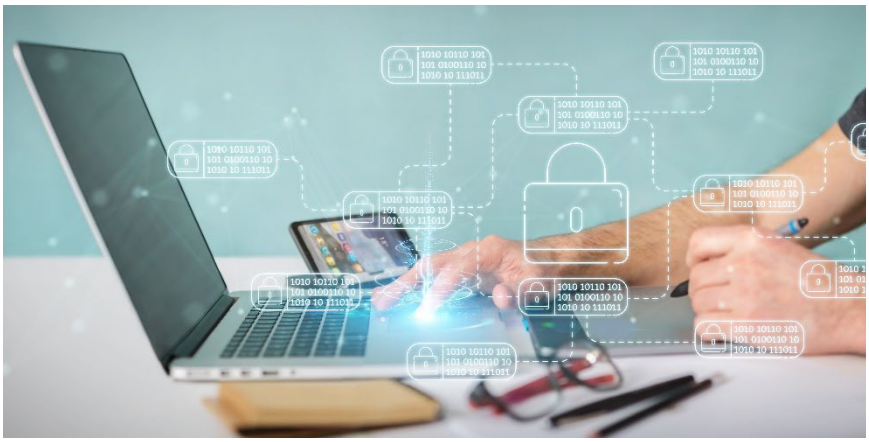
A-SIT

Zentrum für sichere Informationstechnologie - Austria

KÜNSTLICHE INTELLIGENZ UND IT-SICHERHEIT



Künstliche Intelligenz und IT-Sicherheit

Autor:
Lukas Alber
Tel: +43 316 873 - 5559
Mail: lukas.alber@iaik.tugraz.at

Datum: 05.10.2022

Abstract/Zusammenfassung:
Mit fortschreitender, digitaler Transformation und wachsendem Arbeitskräftemangel werden viele Prozesse automatisiert. Reine Softwarelösungen können komplexere Prozesse aber oft nicht hinreichend abbilden. Darum wird auf künstliche Intelligenz gesetzt, um ein höheres Maß an Automatisierung zu erreichen. Anwendungsgebiete reichen von vollautonomen Robotern über künstlerische Gestaltungen bis hin zu IT-Sicherheitsanwendungen. In Letzterem ist künstliche Intelligenz nicht nur Heilsbringer, sondern auch Damoklesschwert, denn sie kann sowohl helfen, Angriffe automatisiert zu verhindern, als auch automatisiert durchzuführen. Hierbei kann die künstliche Intelligenz selbst zum Ziel des Angreifers werden.

Dieser Bericht gibt eine kurze, aber prägnanten Einführung zu künstlicher Intelligenz (KI), ihrer Geschichte und aktuellen Methoden. Anschließend gibt der Bericht einen Überblick über einerseits aktuelle Forschung zum Einsatz von KI in der klassischen IT-Sicherheit sowie andererseits aktuellen Angriffsmethoden auf KI.

Inhalt

1. Einleitung

In den letzten Jahren hat der Bereich der künstlichen Intelligenz einen Frühling erlebt. Fortschritte, wie z.B. im Deep Learning, haben zu einer weit verbreiteten Anwendung von künstlicher Intelligenz (KI) geführt. Der Vorteil liegt hier in der Möglichkeit zur hohen Automatisierung von Prozessen, welche oft mit Schlagworten wie digitale Transformation und Industrie 4.0 in Zusammenhang gebracht werden. Die Anwendungsbereiche im Unternehmensalltag reichen von vollautonomen Robotern[1], welche den Personalaufwand reduzieren, über künstlerische Bildgestaltung [1], bis hin zu IT-Sicherheitsanwendungen [2].

Leider ist KI nicht nur Heilbringer, sondern auch Damoklesschwert. Denn es gibt nicht nur positive Anwendungen von KI, sondern auch kriminelle Elemente können sie nutzen, um schadhafte Ziele zu verfolgen. Darum stellt sich nicht nur die Frage: „Wie kann KI unsere IT-Security Systeme verbessern?", sondern auch „Welchen Attacken ist KI selbst ausgesetzt?". Diesen Fragen wollen wir in diesem Bericht auf den Grund gehen, indem wir auf aktuelle Forschung in diesem Bereich zurückblicken.

Wie mit vielen anderen spannenden Technologien in der Geschichte durchläuft auch künstlichen Intelligenz einen Hype Cycle, in dem sich Jahre der großen Begeisterung mit Jahren marginaler Entwicklung und Förderung abwechseln. Nach zwei KI-Wintern sind wir höchstwahrscheinlich wieder an einem Höhepunkt eines Cycles angelangt. Oft verschwimmen nun sowohl das Verständnis des Möglichen als auch die Definition von künstlicher Intelligenz selbst.

Laut einer Studie nutzen nur 40 % der KI-Startups echte künstliche Intelligenz [3]. Um dies zu verstehen, muss man zwischen Datenanalytik und künstlicher Intelligenz unterscheiden: Ersteres ist ein statischer Analyseprozess über große Datenmengen, um Rückschlüsse zu ziehen. Im Gegensatz zu künstlicher Intelligenz gibt hier keinen selbstlernenden oder iterativen Prozess. Künstliche Intelligenz wird mit zunehmender Datenmenge immer fähiger und autonomer, und kann selbständig Aufgaben erledigen und kognitive Fähigkeiten reproduzieren.

Weiters wird zwischen starker und schwacher KI unterschieden. Schwache künstliche Intelligenz kennen wir bereits von Produkten großer Technologiekonzerne, wie Google und Microsoft, welche mit ihr gerne ihre Produkte aufwerten. Schwache KIs erledigen nur enggefasste Aufgabenbereiche auf menschlichem oder höherem Niveau. Starke KI andererseits ist nicht auf einen einzelnen Use Case beschränkt und kann allgemeine menschliche Intelligenz bzw. sogar ein eigenes Bewusstsein erreichen. Bis jetzt ist noch nicht gelungen, eine starke KI zu erschaffen [2].

Schwache künstlichen Intelligenz, mit welcher wir uns beschäftigen (in Englisch „weak AI"), wird in mehrere Unterstufen eingeteilt: Einerseits Assisted Intelligence, welche bestehende Prozesse verbessert bzw. teilautomatisiert. Eine Stufe höher Augmented Intelligence, welche Aufgabenlösungen ermöglicht, welche sonst nicht möglich wären. Und letztlich Autonomous Intelligence, welche Maschinen ermöglicht, selbstständig zu agieren.

Bereits in den 80er Jahren, vor dem ersten KI-Winter, war künstliche Intelligenz in Form von Expertensystemen das Maß der Dinge. Diese Systeme nutzen eine Wissensdatenbank und ein regelbasiertes Schlussfolgern mithilfe von fuzzy Logic [4], um Denkmuster von Spezialisten zu imitieren. Doch durchsetzen konnten sich diese Systeme nur in Nischen. Die hochgesteckten Ziele konnten nicht annähernd erreicht werden und der Hype flachte ab. Doch Errungenschaften wie Backpropagation beeinflussten die Entwicklung der folgenden Jahre maßgeblich [2].

Maschinelles Lernen (Machine Learning) ist ein Teilbereich des Felds, den der Mantelbegriff künstlichen Intelligenz umschreibt. Machine Learning (ML) beschreibt das Lernen aus Daten, und das Nutzen eines daraus entstehenden Modells, um daraus folgend Schlüsse zu ziehen. Das Lernen ist hier ein andauernder Prozess, der das Verständnis

[1] https://www.energy-robotics.com/post/automating-industrial-inspection-rounds-with-autonomous-mobile-robots, besucht am 12.09.2022

des zu lösenden Problems durch die KI kontinuierlich verbessert. Bevor jedoch eine Maschine mit dem Lernen beginnen kann, müssen die Daten aufbereitet werden. Auf diesen Schritt entfallen in der Praxis meist 80 % der Arbeit [2].

Ein wichtiger Aspekt der Datenaufarbeitung ist Feature Engineering. In diesem Schritt wird Expertenwissen auf den Datensatz angewendet, um die Aussagekraft der Inputfeatures zu verbessern. Gleichzeitig ist das Ziel, die Dimensionalität des Datensatzes zu reduzieren, da sich mit steigender Anzahl an Inputfeatures der Aufwand exponentiell erhöht (Curse of Dimensionality [5]). Dafür kann auch Feature Selection betrieben werden, welche alle wenig aussagenden Inputfeatures entfernt. Dies kann mit Kernel-Methoden, Principal Component Analysis (PCA), oder andere „unsupervised Learning" Techniken erreicht werden.

Der Gegenentwurf dazu ist Deep Learning, welches auf Feature Engineering verzichtet und eher auf Feature Learning setzt. Dabei haben die Entwickelungen der letzten Jahre in Bereich von tiefen neuronalen Netzwerken, dies erst möglich gemacht. Voraussetzungen für diese Entwicklung waren die riesigen Datenmengen (Big Data) und massive Steigerung der Rechenleistung, welche sich in den letzten Jahren ergeben hatten. Beim Feature Learning lernt der Algorithmus selbst, welche Features in der Datenmenge von Bedeutung sind, und erhöht somit seinen Lernfortschritt selbst.

Dieser Bericht über künstliche Intelligenz in der IT-Sicherheit und der Sicherheit von künstlicher Intelligenz wird sich hauptsächlich mit Methoden aus dem Machine Learning auseinandersetzten, da auf diesem Gebiet der Fokus der aktuellen Forschung liegt. Im Kapitel „Vorwissen" werden Grundlagen erklärt, die für das Verständnis von Machine Learning wichtig sind. Im darauffolgenden Kapitel „KI in den Diensten der IT-Sicherheit" berichtet der Text über KIs, welche automatisiert Angriffe aus der klassischen Cybersicherheit durchführen, aber auch die Verteidigung übernehmen können. Schließlich informiert der Bericht in Kapitel „Angriffe auf künstliche Intelligenz" über aktuelle Forschung zur Angreifbarkeit von KI-Systemen selbst.

2. Vorwissen

In diesem Kapitel wird auf Grundlagen über maschinelles Lernen (Machine Learning) eingegangen, welche für das Verständnis der Technologie wichtig sind. Sie können helfen, die weiteren Ausführungen über Sicherheit und KI zu verstehen.

2.1. Lernmethoden

Grundsätzlich kann man Machine Learning Methoden nach der Art und Weise ihres Lernens kategorisieren. Auf die einzelnen Lernmethoden wird in den folgenden Absätzen genauer eingegangen [2]:

- Überwachtes Lernen (supervised learning) lernt von einem Datenset mit gelabelten Daten. Die Maschine lernt daraus Muster, welche sie später in unbekannten Daten sucht, um Schlüsse aus diesen Daten zu ziehen. In der IT-Sicherheit können Methoden aus dem überwachten Lernen dazu verwendet werden, um bereits bekannte Sicherheitsrisiken eindeutig wiederzuerkennen. Typische überwachte Lernalgorithmen sind zum Beispiel lineare Regression, Random Forest und Support Vector Machines (SVMs).

- Unüberwachtes Lernen (unsupervised learning) braucht kein vollständig durchgelabeltes Datenset. Auf die nicht gelabelten Daten können Clustering, Dimensions-Reduktion, oder Anomalie Erkennung angewendet werden. Für die IT-Sicherheit heißt das, dass wir unüberwachtes Lernen für das Erkennen von verdächtigen und noch unbekannten Ereignissen verwenden können. Auch hilft es dabei, den Aufwand für IT-Personal, welches für Sicherheitsüberwachung zuständig ist, zu reduzieren. Typische unüberwachte Lernalgorithmen sind Principal Component Analysis (PCA), K-Means, und Hierachical Clustering.
- Semi-überwachtes Lernen (semi-supervised learning) kommt mit nur einem kleinen Prozentsatz an gelabelten Daten im Datenset aus. Es stellt somit einen Mittelweg aus supervised und unsupervised learning dar. In der IT-Sicherheit ermöglicht es uns darum auch, noch nicht beobachtete Ereignisse zu klassifizieren. Zu dieser Lernmethode gehören Self-Training, Mixture Model und Semi-Supervised Support Vector Machines (SVM).
- Im bestärkenden Lernen (Reinforcement Learning) versucht ein Agent die zu erhaltende Belohnung zu maximieren. Er setzt Aktionen in der Umgebung, in der er sich befindet, und ändert so den Zustand dieser. Die Belohnung wird dann aus der entstandenen Umgebungsänderung ermittelt. Das Ziel ist es, eine Balance zwischen der Exploration neuen Wissens und dem Ausnutzen aus Bekanntem zu erhalten [6]. In der IT-Sicherheit wird bestärkendes Lernen zum Beispiel eingesetzt, um Angreifer, welche das Netzwerk attackieren, zu blockieren und an der Ausbreitung zu hindern. Dabei setzt der intelligente Agent eigenständig Maßnahmen [7].

2.2. Machine Learning Pipeline

Eine ML-Pipeline umfasst alle nötigen Schritte, die für ein ML-Projekt nötig sind, um von rohen Daten zu einer trainierten Maschine, welche erfolgreichen Schlüssen zieht, zu gelangen. Grob unterteilt man in folgende fünf Schritte: Datensammlung, Datenaufbereitung, Feature Engineering, Modellgenerierung und Modellevaluation. Jedoch wird der Schritt Datensammlung oft nicht dazugezählt. Die ersten drei Begriffe wurden bereits in der Einleitung kurz behandelt. Bei der Modellgenerierung wird versucht, eine geeignete Architektur und Hyperparameter für das optimale Lernen zu finden (Hyperparameter dienen der Steuerung des Trainingsalgorithmus). Die Modellgenerierung wird selbst oft (teil)automatisiert [8]. Der Schritt der Modellevaluation bewertet das erlernte Model bzw. die erlernten Modelle und wählt ein passendes aus bzw. kehrt zur Modellgenerierung zurück, falls die Begutachtung negativ ausfällt. Auf weitere Schritte, die in den Bereich MLOps[2] fallen (z.B. Model-Serving) wird in diesem Bericht nicht eingegangen.

[2] https://neu.ro/2021-mlops-platforms-vendor-analysis-report/, besucht am 15.09.2022

3. KI in den Diensten der IT-Sicherheit

Die Fortschritte der letzten Jahre im Machine Learning ermöglichen es, KI auch in der IT-Sicherheit einzusetzen, um den aktiven Schutz von Infrastruktur zu automatisieren und zu verbessern. Des Weiteren kann KI auch präventiv beim Entwickeln und Warten von Software eingesetzt werden, um Sicherheitslücken schon frühzeitig zu erkennen und umfangreiche Sicherheitsüberprüfungen zu automatisieren. Im Folgenden schauen wir uns diese beiden Bereiche genauer an.

Ein Einsatzgebiet ist das automatisierte Planen und Priorisieren der risikoreichsten Probleme. Den chronischen Mangel an IT-Sicherheitsexperten kann dies zwar nicht lösen, aber die Produktivität der Vorhandenen steigern. Ein anderes Einsatzgebiet ist Einbruchserkennung in IT-Systemen. Diese findet z.B. Einsatz in Netzwerken wie Smart Grids [9]. Aktuell wird diese Maßnahme in vielen Projekten ausgerollt, da heutzutage eine ausgiebige Datensammlung in den meisten IT-Projekten zur Verfügung steht.

Einen Schritt weiter geht die aktuelle Forschung, welche versucht, KI zur automatisierten Verteidigung einzusetzen [10], [11]. Diese KI-Systeme können in Echtzeit proaktiv reagieren und den Angreifer in Schach halten, bis ein Sicherheitsexperte eingreifen kann. Das Ziel ist es, den Angreifer vom restlichen System auszusperren bzw. ihn in eine Position zu bringen, in welcher er nur minimalen Schaden anrichten kann [12]. Auch kann die KI den Sicherheitsexperten bei seinem Eingreifen weiterhin unterstützen. Andererseits sind auch Gegenangriffe denkbar, welche den Angreifer außer Gefecht setzen. Dieser Punkt ist auch ein Thema in der Forschung. Man befürchtet ein Wettrüsten zwischen Angreifer- und Verteidigerseite.

Des Weiteren wird auch auf die gezielte Irreführung der Angreifer gesetzt. Ein Beispiel hierfür ist die gezielte Generierung von Honeypots durch die KI, welche den Angreifer in die Irre führen. Dies wurde z.B. in Form von SQL Injection Honey Pots gezeigt [13].

Passend hierzu werden auch angreifende KIs verwendet, um automatisiert Penetrationstests zu generieren und Sicherheitslücken in bestehender Software zu entdecken. Es werden z.B. automatisiert schädliche Payloads generiert, um die Schwächen einer Web Applikation Firewall zu lernen [14]. Dass selbst-lernende Agenten Firewalls erfolgreich umgehen und substanzielle Schäden anrichten können, wurde bereits bewiesen [15].

Gegnerische und verteidigende künstliche Intelligenzen werden in sogenannten Gyms trainiert. Das sind abgeschlossene Umgebungen, in welchen die anzugreifenden bzw. zu verteidigenden IT-Systeme simuliert werden. Hier kann der selbst-lernende Agent ohne Auswirkungen auf das echte System trainiert werden, und das in einer Geschwindigkeit, welche weit über der Echtzeit liegt (begrenzt durch die zur Verfügung stehende Hardware). Der Begriff wurde von OpenAI's Gym Toolkit geprägt [16]. Ein Beispiel für eine mit einem Gym trainierte KI ist WAF-A-MoLE, welche das Ziel verfolgt, SQL Injections an der Firewall vorbeizuschleusen [14].

Im Development-Bereich wird KI, z.B. eingesetzt, um das Schwachstellenmanagement und die Priorisierung zur erleichterten. Beispiele dafür sind kontinuierliche und real-time Risikovorhersagen und ein automatisiertes Risiko-basiertes Schwachstellenmanagement[3]. Auch im Secure Development Lifecycle kann KI zur Teilautomatisierung verschiedener Schritte eingesetzt werden, z.B. teilautomatisierte Testgenerierung[4] und eine kontinuierliche und dynamische Sicherheitsanalyse durch einen KI-Agenten.

[3] https://www.balbix.com/product-overview/, besucht am 20.09.2022
[4] https://github.com/features/copilot, besucht am 20.09.2022

4. Angriffe auf künstliche Intelligenz

Adversarial Machine Learning (AML) bezeichnet den Prozess der Informationsextraktion oder der Manipulation von ML-Systemen, um ein gewünschtes Ergebnis zu erhalten. Dabei werden das Verhalten und die Eigenschaften des ML-Systems extrahiert und der Input ins System manipuliert. Das amerikanische National Institute of Standards and Technology veröffentlichte hierzu bereits einen Draft[5], um für eine einheitliche Taxonomie zu sorgen [17].

Auch Papernot et al. [18] veröffentlichten 2018 ein einheitliches Threat Model für Machine Learning, welches ein strukturiertes Erörtern von Sicherheits- und Privacylücken in diesem Bereich erlaubt. Außerdem präsentieren sie eine vereinheitlichte Taxonomie, um die unterschiedlichen involvierten IT-Fachgebiete, die in der Machine Learning Sicherheit aufeinandertreffen, besser zu verknüpfen [18].

Die Autoren unterscheiden in ihrer Arbeit zwischen dem Lernen und dem Schlussfolgern in einem feindlichen Umfeld. Als Hauptangriffsziel während des Lernvorgangs wurde die Integrität der Daten identifiziert. Bezüglich Privacy und Vertraulichkeit des Datensatzes wird darauf hingewiesen, dass das Lernen auf einer nicht öffentlichen Maschine stattfindet, und darum ein Problem der traditionellen Access Control darstellt, was außerhalb des Rahmens dieses Berichts liegt. Es wird zwischen folgenden Angriffen unterschieden:

- Bei Label-Manipulierung versucht der Angreifer auf die Labels der Datenpunkte einzuwirken, um spätere Schlussfolgerungen zu verfälschen. Biggio et al. [19] konnten zeigen, dass ein Vertauschen der Labels bei 40 % zufällig ausgewählter Daten aus dem Satz zu völlig falschen Schlussfolgerungen führen kann. Außerdem konnte gezeigt werden, dass man mit einem optimal gewählten Subset an Datenpunkten dem Angreifer ermöglicht, mit 10 Prozentpunkte weniger Datenmanipulationen auszukommen. Es ist aber zu beachten, dass je nach Datensatz und der zugrundeliegenden Struktur des Musters, der Prozentsatz an nötigen Daten unterschiedlich ausfallen kann [19]–[21].
- Bei Input-Manipulierung versucht der Angreifer auf die gesamte Information der Datenpunkte einzuwirken, um spätere Schlussfolgerungen zu verfälschen. Diese Art von Data-Poisoning kann auf die Trainingsdaten direkt oder indirekt vor der Datenaufbereitung der Daten stattfinden. Verschiedene Publikation zeigen die hohe Wirksamkeit solcher Attacken [22]–[26].

Deutlich mehr Angriffsmöglichkeiten gibt es im Schlussfolgerungsschritt der ML-Pipeline. Man unterscheidet zwischen White-Box und Black-Box Angreifer. Während Ersterem Informationen über das Innere des ML-Models (Parameter und Architektur) zur Verfügung stehen, müssen Letztere ohne auskommen [18]:

- White-Box Angriffe auf die Integrität des ML-Systems können durch direkte [27]–[29] und indirekte [30], [31] Manipulation (Pipeline-Deformation) der Modellinputs stattfinden [32]. Diese Kategorien werden auch außerhalb der Klassifikation, z.B. im Reinforcement Learning angewendet. White-Box Angriffe auf die Privacy andererseits sind trivial, da das Model zur Verfügung steht.
- Black-Box Angriffe stellen ein realistischeres Szenario für Angreifer dar. Dieser hat meist kein Wissen über da System, kann aber beobachten, wie dieses reagiert. Ähnlich wie in der Kryptografie wird hier gerne ein Oracle als Threat Model verwendet. Da viele ML-Services über eine API als Cloudservice erreichbar sind, erfreut sich das Model wachsender Beliebtheit. Wieder unterscheiden wir bei Angriffen auf die Integrität des Models zwischen direkter und indirekter Manipulation der Inputs. Im Black-Box Szenario sind besonders Angriffe auf die Privacy und Vertraulichkeit der Modelldaten interessant. Membership-Attacken [33], [34], z.B. versuchen zu erkennen, ob gewisse Daten Teil des Trainingssets

[5] https://www.nccoe.nist.gov/ai/adversarial-machine-learning, besucht am 21.09.2022

waren. Bei Trainingsdaten-Extraktion wird, wie der Name schon sagt, versucht, einzelne Datenpunkte auszulesen (z.B. Patientendaten [35]). Des Weiteren gibt es auch noch die Modellextraktion. Hier versucht der Angreifer, geistiges Eigentum zu stehlen, in dem er versucht, das ML-Model mit gezielten Abfragen zu rekonstruieren [36].

5. Fazit

In diesem Bericht wurde einerseits eine Einführung zu künstlicher Intelligenz (KI) gegeben und andererseits ein Überblick über Chancen und Gefahren dieser Technologie in der IT-Sicherheit gegeben. Im ersten Teil, der Einführung, grenzte der Bericht den Begriff KI ab und erklärt die gängige Taxonomie. Auch wurde der Unterschied zu maschinellem Lernen (ML) und KI erklärt. Weiters erläuterte der Bericht die Grundlagen des ML und zusammenhängende Begriffe wie Deep Learning. Dank dieser Grundlagen, wie der Pipeline für maschinelles Lernen, wurden im zweiten Teil gängige Angriffs- und Verteidigungsmuster durch KI, inklusive Angriffe auf die KI selbst, erklärt. Hierzu wurden verschiedenste Publikation aus der aktuellen Wissenschaft herangezogen und zitiert.

Der Bericht soll als Startpunkt dienen, um sich in weiterer Folge mit dem Thema „KI und IT-Sicherheit" auseinander zu setzen. Noch stecken viele der besprochenen Möglichkeiten in den Kinderschuhen, aber in den nächsten Jahren ist mit starken Entwicklungen in diesem Bereich zu rechnen. Und das nicht nur im wissenschaftlichen, sondern auch im privatwirtschaftlichen Bereich. Viele Unternehmen, die auf IT-Sicherheit spezialisiert sind, bieten zunehmend KI-basierte Produkte[6][7][8][9] an.

Referenzen

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CoRR*, vol. abs/2112.10752, 2021, [Online]. Available: https://arxiv.org/abs/2112.10752

[2] P. Wennker and others, "Künstliche Intelligenz in der Praxis," *Anwendung in Unternehmen und Branchen: KI*, 2020.

[3] P. Olson, "Nearly Half Of All 'AI Startups' Are Cashing In On Hype," *Forbes. Com*, 2019.

[4] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965, doi: https://doi.org/10.1016/S0019-9958(65)90241-X.

[5] C. R. Taylor, "Dynamic programming and the curses of dimensionality," in *Applications of dynamic programming to agricultural decision problems*, CRC Press, 2019, pp. 1–10.

[6] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996, doi: 10.1613/jair.301.

[6] https://www.balbix.com/product-overview/, besucht am 21.09.2022
[7] https://winder.ai/, besucht am 21.09.2022
[8] https://www.juniper.net/us/en/solutions/artificial-intelligence-for-it-operations-aiops.html, besucht am 21.09.2022
[9] https://www.crowdstrike.com/, besucht am 21.09.2022

[7] Y. Han *et al.*, “Reinforcement Learning for Autonomous Defence in Software-Defined Networking,” *CoRR*, vol. abs/1808.05770, 2018, [Online]. Available: http://arxiv.org/abs/1808.05770

[8] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, “Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms,” in *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, 2013, pp. 847–855. doi: 10.1145/2487575.2487629.

[9] M. N. Kurt, O. E. Ogundijo, C. Li, and X. Wang, “Online Cyber-Attack Detection in Smart Grid: A Reinforcement Learning Approach,” *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5174–5185, 2019, doi: 10.1109/TSG.2018.2878570.

[10] Y. Han *et al.*, “Reinforcement Learning for Autonomous Defence in Software-Defined Networking,” in *Decision and Game Theory for Security - 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29-31, 2018, Proceedings*, 2018, vol. 11199, pp. 145–165. doi: 10.1007/978-3-030-01554-1_9.

[11] R. Elderman, L. J. J. Pater, A. S. Thie, M. M. Drugan, and M. A. Wiering, “Adversarial Reinforcement Learning in a Cyber Security Simulation,” in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence, ICAART 2017, Volume 2, Porto, Portugal, February 24-26, 2017*, 2017, pp. 559–566. doi: 10.5220/0006197105590566.

[12] M. Panfili, A. Giuseppi, A. Fiaschetti, H. B. Al-Jibreen, A. Pietrabissa, and F. D. Priscoli, “A Game-Theoretical Approach to Cyber-Security of Critical Infrastructures Based on Multi-Agent Reinforcement Learning,” in *26th Mediterranean Conference on Control and Automation, MED 2018, Zadar, Croatia, June 19-22, 2018*, 2018, pp. 460–465. doi: 10.1109/MED.2018.8442695.

[13] L. Erdödi, Å. Å. Sommervoll, and F. M. Zennaro, “Simulating SQL injection vulnerability exploitation using Q-learning reinforcement learning agents,” *J. Inf. Secur. Appl.*, vol. 61, p. 102903, 2021, doi: 10.1016/j.jisa.2021.102903.

[14] L. Demetrio, A. Valenza, G. Costa, and G. Lagorio, “WAF-A-MoLE: evading web application firewalls through adversarial machine learning,” in *SAC ’20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 - April 3, 2020*, 2020, pp. 1745–1752. doi: 10.1145/3341105.3373962.

[15] K. Hammar and R. Stadler, “Finding Effective Security Strategies through Reinforcement Learning and Self-Play,” in *16th International Conference on Network and Service Management, CNSM 2020, Izmir, Turkey, November 2-6, 2020*, 2020, pp. 1–9. doi: 10.23919/CNSM50824.2020.9269092.

[16] G. Brockman *et al.*, “OpenAI Gym,” *CoRR*, vol. abs/1606.01540, 2016, [Online]. Available: http://arxiv.org/abs/1606.01540

[17] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, “A taxonomy and terminology of adversarial machine learning,” *NIST IR 8269*, pp. 1–29, 2019.

[18] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “SoK: Security and Privacy in Machine Learning.,” *Proceedings - 3rd IEEE European Symposium on Security and Privacy, EURO S and P 2018*, pp. 399–414, Jul. 2018, doi: 10.1109/EuroSP.2018.00035.

[19] B. Biggio, B. Nelson, and P. Laskov, “Support vector machines under adversarial label noise,” in *Asian conference on machine learning*, 2011, pp. 97–112.

[20] H. Xiao, H. Xiao, and C. Eckert, “Adversarial label flips attack on support vector machines,” in *ECAI 2012*, IOS Press, 2012, pp. 870–875.

[21] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE J Biomed Health Inform*, vol. 19, no. 6, pp. 1893–1905, 2014.

[22] M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 405–412.

[23] B. Biggio *et al.*, "Poisoning behavioral malware clustering," in *Proceedings of the 2014 workshop on artificial intelligent and security workshop*, 2014, pp. 27–36.

[24] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," 2015.

[25] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.

[26] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2017, pp. 262–275.

[27] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[28] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 2574–2582. doi: 10.1109/CVPR.2016.282.

[29] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006, Taipei, Taiwan, March 21-24, 2006*, 2006, pp. 16–25. doi: 10.1145/1128817.1128824.

[30] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017. [Online]. Available: https://openreview.net/forum?id=HJGU3Rodl

[31] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, 2016, pp. 1528–1540. doi: 10.1145/2976749.2978392.

[32] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, 2016, pp. 372–387. doi: 10.1109/EuroSP.2016.36.

[33] J. Hayes, L. Melis, G. Danezis, and E. de Cristofaro, "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks," *CoRR*, vol. abs/1705.07663, 2017, [Online]. Available: http://arxiv.org/abs/1705.07663

[34] J. Ye, A. Maddi, S. K. Murakonda, and R. Shokri, "Enhanced Membership Inference Attacks against Machine Learning Models," *CoRR*, vol. abs/2111.09679, 2021, [Online]. Available: https://arxiv.org/abs/2111.09679

[35] M. Fredrikson, E. Lantz, S. Jha, S. M. Lin, D. Page, and T. Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," in *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014*, 2014, pp. 17–32. [Online]. Available:

https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew

[36] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing Machine Learning Models via Prediction APIs,” in *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016*, 2016, pp. 601–618. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer