SICHERHEIT VON UND DURCH MASCHINELLES LERNEN

Impulspapier | Dezember 2020

Zusammenfassung

Das vorliegende Impulspapier adressiert die Schnittstellen von Sicherheit im Sinne von Cybersecurity und maschinellem Lernen aus mehreren Perspektiven: Zum einen wird die Sicherheit bei der Nutzung von Verfahren des maschinellen Lernens betrachtet. Zum anderen wird aber auch maschinelles Lernen als Mechanismus zum Herstellen von IT-Sicherheit betrachtet. Dabei werden Aspekte herausgegriffen, die aus unserer Sicht von besonderer und aktueller Bedeutung sind: Die Verbesserung der Widerstandsfähigkeit gegen gezielte, teilweise neuartige Angriffe wird im Abschnitt „Härtung“ betrachtet. Herausforderungen hinsichtlich der Privatheit widmet sich der Abschnitt „Datenschutz“. Da es oft schwer ist, Nutzen und Erfolgschancen beim Einsatz von Maschinellem Lernen abzuschätzen, liefert der Abschnitt „Performanz und Vergleichbarkeit“ hierzu Anregungen. Die Ergebnisse, die maschinelles Lernen liefert, werden von Nutzern oft als schwer nachvollziehbar angesehen. Im Abschnitt „Interpretierbarkeit“ finden sich Strategien, ein besseres Verständnis von maschinellem Lernen zu erreichen. Die unerkannte Nutzung von maschinellem Lernen, beispielsweise bei Deep Fakes, wird vielfach als Bedrohung wahrgenommen. Gegenmaßnahmen widmet sich der Abschnitt „Nachvollziehbarkeit“.

Jedem der Abschnitte sind Handlungsempfehlungen an Politik, Behörden, Unternehmen und Wissenschaft nachgestellt, die dazu beitragen können, dass Sicherheit und maschinelles Lernen aufeinander abgestimmt umgesetzt werden können. Die Politik sollte Anreize schaffen, die Chancen durch maschinelles Lernen nicht ohne eine umfassende Betrachtung der Sicherheitsaspekte zu realisieren. Parallel dazu sollten Unternehmen frühzeitig erkennen, welche Risiken beim Einsatz von maschinellem Lernen entstehen können und wie diesen entgegengewirkt werden kann. Diese Prozesse sollten Behörden durch geeignete Rahmenbedingungen und beispielhafte Umsetzungen unterstützen. Der Wissenschaft fällt die Aufgabe zu, bei der rasanten Entwicklung des maschinellen Lernens nicht die Sicherheitsaspekte aus den Augen zu verlieren und frühzeitig über Gefahren, Schwächen und Gegenmaßnahmen zu informieren.

1 Ausgangslage

Während sich Maschinelles Lernen (ML) und die IT-Sicherheit jede für sich rasant entwickeln, wurden insbesondere in den letzten Jahren vielversprechende Synergien zwischen diesen beiden Gebieten entdeckt. So lassen sich einerseits mit den Methoden des ML oftmals schwierige Probleme der IT-Sicherheit nicht nur angehen, sondern teilweise auch lösen. Andererseits ist die Perspektive der IT-Sicherheit von großem Nutzen, um ML-Methoden zu betrachten, zu durchleuchten, zu testen und abzusichern. Diese Symbiose ist nicht nur ein Gewinn für beide Gebiete für sich betrachtet – es entsteht vielmehr ein neues Forschungsfeld, das als zukünftige Schlüsseltechnologie betrachtet werden sollte.

Spricht man von ML und IT-Sicherheit, ist die verbreitetste Perspektive die, dass ein Computer bei der Erkennung und Bekämpfung von IT-Sicherheitsvorfällen helfen soll. ML kann bei der IT-Sicherheit besonders dann helfen, wenn die zu untersuchenden Daten umfangreich und unstrukturiert sind. Hier versagen Ansätze, die auf starren Regeln basieren. ML kann anhand von Beispielen lernen, wie Angriffe aufgebaut sind und wie der Normalfall aussehen sollte. Sie kann aber auch dabei helfen, große Datenmengen initial zu strukturieren, um einen schnelleren Überblick über Vorfälle zu erhalten.

Viele weitere Anwendungen sind nicht nur Gegenstand aktueller Forschung, sondern werden derzeit aktiv im Markt etabliert. Diese reichen von der Erkennung von Auffälligkeiten in komplexen Informationssystemen über das automatisierte, KI-gestützte Auffinden von Schwachstellen in Software bis hin zu Sicherheitslösungen der Usable Security, welche den Menschen und seinen Umgang mit Sicherheitstechnologien ins Zentrum rücken.

Angriffserkennung und -analyse
In der IT-Sicherheit ist der Einsatz von ML bereits heute in vielen Bereichen der Angriffserkennung und -analyse etabliert. Die Erkennung von Spam-Mails ist ein bekanntes Beispiel dafür, dass ML-Verfahren dafür eingesetzt werden, gewünschte von unerwünschten E-Mails zu unterscheiden. Autorenzuordnung kombiniert Forensik und auf ML-gestützte Verarbeitung von Texten. Deep Learning hilft bei der Identifizierung illegaler Bilder und hat die Erkennung von Schadsoftware und von Angriffen auf Netzwerke verbessert.

Durch die Möglichkeit, große und auch heterogene Datenmengen automatisiert auszuwerten, ergeben sich völlig neue Möglichkeiten der Lagebilderstellung. Neben der reinen Visualisierung etablieren sich derzeit Lösungen, welche auf die vollautomatisierte Auswertung großer Mengen von Metadaten zielen, um durch Frühwarnsysteme IT sicher betreiben zu können. Noch in frühen Phasen der Entwicklung sind Ansätze der Predictive Security, die Gefahren im Vorfeld erkennen sollen, um frühzeitig entsprechende Gegenmaßnahmen präventiv einleiten zu können. Auch bei der Analyse und Forensik im Falle eines bereits erfolgten Angriffes kommen zunehmend KI-Methoden zum Einsatz.

Beispiel: Spam-Erkennung
Google setzt unter anderem sein Framework TensorFlow zur Spam-Erkennung im hauseigenen E-Mail-Dienst GMail ein.[1] [2] Dabei ist es Teil einer Lösung, die sowohl regelbasierte Ansätze als auch Methoden des ML miteinander kombiniert. Die Aufgaben sind dabei geteilt: Regelbasierte Ansätze filtern bekannte Spam-Muster, während maschinelles Lernen zum Erkennen von neuen Spam-Typen verwendet wird. Nach eigenen Angaben werden 99,9 % aller Spam-E-Mails hiermit erkannt. Da Spam-Erkennung bereits frühzeitig ML eingesetzt hat, finden sich hier viele der Standard-Methoden des ML wieder.

Beispiel: Erkennung von Kreditkartenbetrug
Bei der Erkennung von Kreditkartenbetrug ist die Herausforderung, in einer großen Menge von Transaktionen Betrugsfälle zu erkennen, ohne dabei den Zahlungsverkehr zu verzögern. Gleichzeitig muss die Wahrscheinlichkeit von Fehlalarmen minimal sein. Dazu werden unterschiedliche Varianten des ML eingesetzt.[3] [4] Unsupervised Learning wird verwendet, um Muster zu erkennen und Bezüge herzustellen. Mit Supervised Learning werden Bezüge zwischen bekannten Betrugsfällen und neuen Fällen hergestellt.

Automatisieren von Angriffen – Einsatz von ML zur Härtung und zum Testen
Zur Erkennung von Schwachstellen von Lösungen, die unter anderem auch selbst auf ML basieren können, ist ML ebenfalls ein wertvolles Werkzeug. So können Lücken hinsichtlich des Datenschutzes eines trainierten Netzes durch den Einsatz verschiedener Methoden des ML entdeckt werden. Natürlich kann ML aber auch dazu eingesetzt werden, Angriffe auf IT-Systeme oder Nutzer zu automatisieren, beispielsweise, um skalierbare Spear-Phishing-Angriffe auf Personen anhand von Social-Media-Profilen durchzuführen. Ganz allgemein wird auch Hacking inzwischen durch ML-Methoden unterstützt, was sich insbesondere beim systematischen Testen auf Schwachstellen anbietet. In das Testen von Software haben KI-Methoden ebenfalls bereits Einzug gehalten. Dies umfasst sowohl die statische Quellcodeanalyse als auch das randomisierte Testen (Fuzzing) von Binärdateien und eingebetteter Software.

Beispiel: Phishing
Ein Feld, in dem ML bereits verbreitet eingesetzt wird, ist Phishing[5]. Da es hier darum geht, möglichst gut auf das Ziel abgestimmte Mitteilungen zu erstellen, hilft ML in Kombination mit Open Source Intelligence, geeignete Informationen zu sichten und darauf basierend gezielte Angriffe (Spear-Phishing) durchzuführen.

Beispiel: Malware
Von IBM wurde bereits 2018 mit DeepLocker[6] gezeigt, dass ML, in diesem Fall Deep Learning, Malware so verschleiern kann, dass sie schwerer detektierbar wird. Durch ML wurden Eigenschaften der Malware optimiert, wodurch sowohl die Konditionen, unter denen sie aktiv wird, als auch der eigentliche Schadcode besser verborgen wurden.

Managed Security
ML wird heute verbreitet für die Erkennung von Angriffen auf Netzwerke eingesetzt. Intrusion Detection Systems (IDS) und Security Information und Event Management (SIEM) nutzen die Möglichkeiten von ML bei der Erkennung von Anomalien im Netzwerk durch vielfältige Sensordaten. Werden Vorfälle in einem Netzwerk nicht nur auf Basis von Daten des Netzes bewertet, sondern auch anhand von Informationen aus aktuellen Angriffskampagnen oder dem Überwachen von Internetforen, wächst die Komplexität der Systeme weiter. Entsprechende Lösungen werden von Unternehmen angeboten, die sich auf die Detektion auch komplexer Angriffe spezialisiert haben. ML wird auch zunehmend eingesetzt, um die Kommunikation zwischen potentiellen Angreifern, die sich beispielsweise in Foren austauschen, zu erkennen und hinsichtlich der damit verbundenen Risiken zu bewerten. Hier ist die Verarbeitung unstrukturierter Textdaten eine etablierte Disziplin, die wiederum zahlreiche Mechanismen des ML verwendet.

Beispiel: SIEM und Congnitive Computing
Um die Security-and-Event-Management-Lösung um analytische Komponenten zu erweitern, nutzt IBM die Plattform Watson[7], die Methoden des Cognitive Computing realisiert. Mit ihr werden die Indikatoren der SIEM-Lösung interpretiert, um mehr über die Natur eines Angriffs zu erfahren. Sie kann beispielweise Experten eines SOC dabei unterstützen, laufende Angriffe zu bewerten.

2 Härtung von ML

IT-Sicherheit steht in der Praxis vor der Herausforderung, dass Methoden zum Erreichen von Schutz durch neu entwickelte Angriffe umgangen werden. Es handelt sich um ein kontinuierliches Agieren und Reagieren. Dies gilt auch für den Bereich ML und Sicherheit. Die vorhergehenden Beispiele zeigen, dass ML in einigen Bereichen bedeutende Verbesserungen beim Herstellen von Sicherheit ermöglicht hat. Naturgemäß hat auch dies Gegenentwicklungen mit sich gebracht. Sowohl in der Forschung als auch in der Praxis sind inzwischen zahlreiche Angriffe auf ML-Lösungen bekannt, die Schwachstellen ausnutzen oder aufzeigen.

Ein Beispiel ist hier der Wettlauf im Bereich der Spam-Erkennung. Wurden als erste Maßnahmen gegen die Erkennung von Verbreitern noch Schlüsselworte verschleiert, indem Buchstaben durch ähnlich aussehende Symbole ersetzt wurden, werden inzwischen auch die sich immer wieder nachtrainierenden Netze angegriffen, in denen zuerst unverfängliche, den später geplanten Spam-Kampagnen ähnelnde Nachrichten verbreitet werden, um auf diese Weise unter dem Schwellwert der Erkennung zu bleiben[8].

Allgemein können Angriffe auf ML sowohl in der Trainingsphase als auch in der Anwendungsphase erfolgen. Eine ML-Lösung kann dabei durchaus anfällig gegen Angriffe in beiden Phasen sein. So sind bei der Erkennung von Verkehrsschildern beim Autonomen Fahren Angriffe bekannt, die die Trainingsphase so verändern, dass einfache Manipulationen an Schildern zu einer fälschlichen Klassifizierung führen[9]. Ebenso können aber Angreifer auch die von einer Kamera aufgenommenen Bilder vor der Klassifizierung so verändern, dass eine fehlerhafte Erkennung erfolgt[10]. Selbst Modifikationen in der realen Welt sind möglich, um eine Klassifizierung zu verwirren[11].

Handlungsempfehlungen
Die Erfahrung der IT-Sicherheitsforschung und auch die Praxis zeigen, dass ein Sicherheitsrisiko nie völlig ausgeschlossen werden kann und sich Angreifer und Verteidiger in einem ständigen Wettlauf befinden. Dementsprechend sind Maßnahmen zum Steigern der Sicherheit von ML immer als eine kontinuierliche Aufgabe zu sehen.

Um eine hohe Sicherheit nach dem aktuellen Stand der Technik zu gewährleisten, sind organisatorische Ansätze zu erwägen. So können Mindestanforderungen an ML-Lösungen gestellt werden, die Sicherheitsaufgaben übernehmen. Dazu zählen normierte Testdatensätze oder Testabläufe, die auch bekannte Angriffe beinhalten und so prüfen, ob ML-Lösungen zumindest bekannte Gegenmaßnahmen umsetzen.

Als Handreichungen sind „Best Practise"-Beispiele zu definieren, die zeigen, wie eine Härtung gegen Angriffe erfolgen kann, wenn hier bereits Lösungen bekannt sind. Da der Einsatz von ML in immer mehr Anwendungen Einzug hält, ist es notwendig, Werkzeuge zum Prüfen und Testen der Sicherheit von ML zu erstellen, damit Entwickler von Software eine effiziente Möglichkeit zur Härtung ihrer Lösungen zur Hand haben.

Bisher sind Anwendungen von ML der dominierende Gegenstand der Forschung; potenziellen Nutzern sollen umgehend Methoden des ML zum Lösen von Problemen zur Verfügung gestellt werden. Gerade die absehbar starke Verbreitung von ML und trainierten Netzen erfordert aber, auch parallel signifikante Forschungsaktivitäten zu den Fragen der Sicherheitsrisiken von ML durchzuführen. Analog zur Krypto-Forschung, wo die Forschung gleichermaßen auf die Konstruktion und das Beweisen bzw. Brechen von Verfahren abhebt, sollten hier Anreize zur kritischen Analyse und Absicherung von ML-Lösungen und deren Sicherheit geschaffen werden.

Daraus leiten sich die folgenden Handlungsempfehlungen an unterschiedliche Kreise ab:

- Politik:
  - Strategisches Ziel wie „Deutschland als Leitmarkt für ML-basierte Sicherheit" setzen, Maßnahmen hierfür entwickeln (wie einschlägige Vergabevorgaben für die öffentliche Hand), sichtbar pushen und umsetzen.
  - Innovationen in ML-basierter Sicherheit fördern, beispielsweise durch das Pushen eines einschlägigen Start-up-Ökosystems, und die Verbesserung der Rahmenbedingungen für einschlägige Forschungsaktivitäten der etablierten Wirtschaft und Industrie.
  - In der Kommunikation Querbezüge zu anderen Themen herstellen, denen ML-basierte Sicherheit nützt, wie beispielsweise Daseinsvorsorge, Digitale Souveränität und sichere digitale Infrastrukturen.
  - Definieren von Anforderungen an die Sicherheit von ML, abhängig von ihrem Einsatzgebiet, um Rechtssicherheit zu schaffen. So kann für Bereiche mit hohem

Risiko, beispielsweise in der Medizin, eine besondere Zertifizierung von Methoden gefordert werden.
  - Rahmenbedingungen für die Aus- und Weiterbildung im Themengebiet verbessern und einschlägige Anreize schaffen, beispielsweise durch gezielte Förderprogramme.
- Behörden:
  - Orientierung geben, beispielsweise durch das Zusammentragen von erprobten Umsetzungsvorschlägen zum sicheren Einsatz von ML.
  - Benchmarking vorantreiben, beispielsweise durch das Erstellen von Testdatensätzen zum algorithmenübergreifenden Vergleich von Resistenzen gegen bekannte Angriffe auf Trainingsdatensätze.
- Unternehmen:
  - Erarbeiten von Risikoabschätzungen bei den Anwendungen des ML und Ableiten einer angepassten Sicherheitsstrategie.
  - Einbinden von Sicherheitsexperten bei der Einführung und dem Betrieb von ML. Dies macht eine Ausbildung entsprechender Experten notwendig, da diese auf dem Arbeitsmarkt nur begrenzt verfügbar sind.
- Wissenschaft:
  - Vertiefen der Forschung hinsichtlich der Sicherheitseigenschaften von Algorithmen des Maschinellen Lernens, beispielsweise bezüglich der Aspekte Robustheit, IT-Sicherheit, Verlässlichkeit, Integrität, Transparenz, Erklärbarkeit, Interpretierbarkeit und Nichtdiskriminierung.
  - Orte mit guten Forschungsbedingungen und Anreize schaffen, in denen neue Methoden des Maschinellen Lernens von Forschergruppen hinsichtlich ihrer Sicherheit evaluiert werden können.

3 Datenschutz und ML

Der Datenschutz kann von hoher Bedeutung im Kontext von ML sein – es geht nicht nur um den gesetzeskonformen Einsatz, sondern auch um die Akzeptanz. Arbeitet ML mit personenbezogenen Daten oder gibt sie diese als Ergebnis aus, können Mechanismen zur Anonymisierung oder Pseudonymisierung notwendig werden. Das ist besonders dann eine Herausforderung, wenn das Training personenbezogene Daten erfordert, für diese Daten aber noch keine Methoden zur zuverlässigen Anonymisierung und Pseudonymisierung bekannt sind. Denn es sind inzwischen Angriffe bekannt, die darauf abzielen, personenbezogene Daten wieder aus einem trainierten Netz zu extrahieren. Dabei können Daten entweder durch das Messen der Reaktionen eines trainierten Netzes iterativ synthetisiert (model inversion) oder ihre Verwendung beim Training nachgewiesen (membership inference) werden.

Darüber hinaus muss auch eine weitere Perspektive betrachtet werden: Durch ML werden Verfahren ermöglicht, die zunehmend invasiv hinsichtlich von weiteren Persönlichkeitsrechten sein können. Ein besonders drastisches Beispiel ist der Fall eines Internetforums, in dem Nutzer ein Verfahren zur Gesichtserkennung auf Basis von ML mit einer Suchmaschine für ein soziales Netzwerk kombinierten und damit Darstellerinnen in erotischen Filmen erkannten und denunzierten[12].

Handlungsempfehlungen

Der Umstand, dass durch ML immer komplexere Verbindungen zwischen Daten hergestellt werden können, erfordert eine Steigerung der Anforderungen an den Schutz personenbezogener Daten. ML verringert den Aufwand, Anonymität und Pseudonymität aufzuheben, was mit einem Stärken der Algorithmen zum Gewährleisten des Datenschutzes zu begegnen ist. Hinsichtlich des Datenschutzes empfiehlt es sich, ML in seiner produktiven Umgebung zu betrachten und Methoden des Privacy-by-Design anzuwenden. Diese erfordern neben technischen Sicherheitsmechanismen auch die Protokollierung von Zugriffen und Möglichkeiten zur Datenkorrektur.

Um personenbezogene Daten in trainierten Netzen zu schützen, können Methoden eingesetzt werden, die auch beim Schutz von Datenbanken verwendet werden. Diese setzen sich primär aus Zusammenführen oder Verrauschen von Daten zusammen, bekannte Ansätze sind hier Differential Privacy sowie k-Anonymity und seine Nachfolger. Das Federated Learning, bei dem nicht die verteilten Daten, sondern lediglich die auf den verteilten Daten trainierten Modelle an eine zentrale Stelle gesendet werden, stellt hier einen weiteren, vielversprechenden Ansatz dar. Sollen die Daten gegen ein Ausspähen durch einen Anbieter von Rechnerinfrastrukturen geschützt werden, ist hier die Forschung an ML-Analysen direkt auf verschlüsselten Daten zu nennen.

Die Herausforderung, personenbezogene Daten vor einer unzulässigen Nutzung durch Dritte zu schützen, ist allerdings nicht alleine technisch lösbar. Sollen Daten in einem sozialen Netzwerk frei verfügbar sein, so kann auch eine Maschine auf diese zugreifen und aus ihnen lernen. Hier hilft nur eine Kombination aus Aufklärung, Gesetzgebung und technischen Hürden.

Daraus leiten sich die folgenden Handlungsempfehlungen an unterschiedliche Kreise ab:

- Politik:
  - Aufstellen verbindlicher Vorgaben hinsichtlich der Nutzung frei zugänglicher personenbezogener Daten unter Berücksichtigung der potentiellen Risiken durch ML.
  - Schaffen von Anreizen zur Entwicklung von ML-Lösungen nach europäischen Standards. Dazu gehören die Entwicklung von Methoden, aber auch das Schaffen von Datengrundlagen durch europäische Alternativen zu großen Datenplattformen.
  - Rechtssicherheit schaffen: Den Gebrauch von Anonymisierungs- und Pseudonymisierungsverfahren bei ML rechtlich, z. B. bezüglich Schutzniveau, einordnen.
- Behörden:
  - Beispielhafte Umsetzungen von ML und Datenschutz, u. a. durch die Berücksichtigung von Mechanismen des Privacy-by-Design in eigenen ML-Anwendungen.
  - Zielgerichtete Vergabepraxis pro ML mit Datenschutz.
- Unternehmen:
  - Frühzeitiges Einbinden von Mechanismen des Datenschutzes in Projekte, die ML und personenbezogene Daten vereinen.
  - Prüfen der Umsetzbarkeit von Projekten auf Basis anonymisierter/pseudonymisierter Daten zur Vermeidung von Datenschutzproblemen.
- Wissenschaft:
  - Erforschen von Methoden zur Herstellung von Privatheit unter Berücksichtigung weit verbreiteter frei verfügbarer personenbezogener Daten.
  - Erforschen der Vereinbarkeit von Privacy-by-Design und effizientem ML.

4 Performanz und Vergleichbarkeit

Eine fundamentale Eigenschaft von ML ist, dass die Ergebnisse immer eine Wahrscheinlichkeit darstellen. Diese Wahrscheinlichkeit beschreibt die Übereinstimmung, die das trainierte System zwischen den vorliegenden Daten und den Trainingsdaten herleiten kann. Dies führt aber auch immanent eine Chance für einen Irrtum mit sich. Je nach Methode und Daten sind hier Fehlerwahrscheinlichkeiten im Bereich von einem Promille bis hin zu 20 Prozent und mehr zu beobachten. Diese Fehlerwahrscheinlichkeiten sind oft unterschiedlich bezüglich des Umstands, ob fälschlich eine positive oder eine negative Übereinstimmung angenommen wird. Verbunden sind beide Eigenschaften allerdings trotzdem: Wenn ein Schwellwert über Zustimmung oder Ablehnung entscheidet, sind sowohl falsche Zustimmung und Ablehnung direkt von ihm abhängig. Vereinfacht kann festgehalten werden, dass ein System, welches seinen Anteil an falschen Ablehnungen mittels des Absenkens des Schwellwerts erreichen will, dies nur mit einem Zuwachs an falschen Zustimmungen erreichen kann. Dies gilt zumindest immer dann, wenn durch ML eine Ja/Nein-Entscheidung erfolgen soll, was allerdings häufig der Fall ist. Wird durch die Verfahren nur ein Grad Übereinstimmung gemessen, so entfällt die finale Entscheidung, die Fehler führen hier zu ungenauen Einschätzungen.

Das Maß, in dem solche Fehler akzeptabel sind, ist stark abhängig von der Anwendung. Während im Marketing eine Fehlerrate von 20 Prozent immer noch ein erfolgreiches Instrument beschreiben kann, ist dies in einer Sicherheitsanwendung eventuell ein Ausschluss-Kriterium: Ist jeder fünfte Empfänger eines Werbeschreibens nicht an einem Produkt interessiert, kann die Kampagne durchaus erfolgreich sein. Wird jede fünfte Transaktion eines Kreditinstituts als Betrugsversuch angesehen und gestoppt oder jede fünfte Internetverbindung unterbunden, da die KI eines Browsers einen Angriff vermutet, ist dies für die Betroffenen unzumutbar. Sind entsprechende Fehlerraten unvermeidbar, sollte eine KI nicht als eine autonome Entscheidungsinstanz angesehen werden, sondern kann als Vorfilter dienen, der einem menschlichen Entscheider substantielle Erleichterung bei der Arbeit verschafft.

Handlungsempfehlungen

Für Anwender von ML ist es bisher schwer abschätzbar, wie nützlich und zuverlässig der Einsatz einer entsprechenden Lösung wirklich ist und wie sich Trainingsaufwand und Performanz zueinander verhalten. Maßzahlen, die neben technischen abstrakten Eigenschaften, die von Konfusionsmatrizen abgeleitet sind, auch Anwendungsparameter einbeziehen, beispielsweise die erwartete Verteilung von positiven und negativen Fällen, sind hier notwendig. Demensprechend ist hier eine interdisziplinäre Forschung erforderlich, die Technik und Nutzersicht harmonisieren.

Notwendig sind hier Strategien zur Aufklärung und zur Bereitstellung einheitlicher Testverfahren. Für Standardaufgaben könnten beispielsweise normierte Testdatensätze entwickelt werden, die eine Vergleichbarkeit zwischen verschiedenen Verfahren erleichtern. Wichtig ist hier eine Herangehensweise, die ein übermäßiges Optimieren von Verfahren auf die Testdaten verhindert, beispielsweise durch eine hybride Prüfung mit einem öffentlich verfügbaren und einem internen Testdatensatz. Für die Beschreibung von Lösungen, die ML in kritischen Aufgaben einsetzen, sollte eine einheitliche Angabe von Eigenschaften gefordert werden, beispielsweise die vollständige Konfusionsmatrix mit einer vorgegeben realitätsnahen Mischung von positiven und ne

gativen Beispielen oder die Ergebnisse der Bearbeitung des vorher genannten öffentlich verfügbaren Testdatensatzes, wobei hier wieder ein Risiko von Überanpassung besteht.

Daraus leiten sich die folgenden Handlungsempfehlungen an unterschiedliche Kreise ab:

- Politik:
  - Hinwirken auf die Einführung verbindlicher Kennzeichnungen der Eigenschaften von ML-Verfahren und standardisierter Testverfahren für ML in kritischen Anwendungsfällen wie Medizin oder öffentlicher Sicherheit.
- Behörden:
  - Bereitstellen von normierten Testdatensätzen, an denen Standardaufgaben des ML durchgeführt werden können und die realistische Szenarien widerspiegeln, beispielsweise hinsichtlich positiver und negativer Fälle.
- Unternehmen:
  - Entwickeln von Anforderungskatalogen für den Einsatz von ML und Erarbeiten der Datengrundlagen, auf denen ML angewandt werden soll. Mit diesen Informationen ist eine bessere initiale Beurteilung der Erfolgschancen von zuverlässigem ML möglich.
- Wissenschaft:
  - Verständigen auf einheitliche Kennzahlen von Algorithmen zur besseren Vergleichbarkeit in wissenschaftlichen Publikationen und eine genaue Beschreibung der Trainings- und Testdaten.

5 Interpretierbarkeit von Ergebnissen

Die Ergebnisse, die Verfahren des ML liefern, sind in vielen Fällen schwer interpretierbar. Eine Klassifizierung erfolgt gegebenenfalls mit einem Konfidenzwert, aber üblicherweise ohne eine Erklärung, wie dieser erreicht wurde. Dies ist primär der schwer erfassbaren Struktur der trainierten Netze, insbesondere bei Ansätzen wie Deep Learning geschuldet. Zwar ist bekannt, wie die Daten vom Eingang über die Ebenen zum Ausgang bzw. Ergebnis gelangen, die tatsächliche Bedeutung der einzelnen Knoten und Kanten auf diesem Weg bleibt aber verborgen.

Solange diese Ergebnisse ausreichen, um Entscheidungen zu treffen, ist dies unproblematisch. Das gilt für eine Vielzahl von Lösungen, in denen ML bei der Nutzung großer Datenmengen hilft und ohne ernsthafte Alternativen ist. Eine Empfehlung auf einem Online-Portal, basierend auf dem eigenen Kaufverhalten und einer Korrelation von gesammelten Daten anderer Kunden, kann gegebenenfalls auch zu einem falschen Vorschlag führen, dieser ist aber in der Regel ohne signifikante Konsequenzen. Treffer führen potentiell zu einem Kauf, Fehler aber nicht zu einem Verlust des Kunden.

Werden entsprechende Verfahren aber beispielsweise in der IT-Forensik eingesetzt, ist eine Interpretierbarkeit deutlich bedeutsamer: Eine Entscheidung eines ML-Ansatzes, welcher als Indiz beispielsweise die Autorschaft eines Bekennerschreibens belegen soll, muss für die im Prozess Beteiligten nachvollziehbar sein.

Aber auch allgemein ist die Interpretierbarkeit wertvoll, da sie dabei hilft, Fehlverhalten von ML-Systemen schneller zu beheben: Wenn schnell erkannt wird, warum beispielsweise Bilder falsch klassifiziert werden, indem durch eine Heatmap die ausschlaggebenden Bildbereiche unterlegt werden, werden oft fundamentale Fehler in den Trainigsdaten aufgedeckt und können beseitigt werden.

Demensprechend ist es ratsam, allgemein an der besseren Interpretierbarkeit von ML-Systemen zu arbeiten. Dies kann durch ein Ableiten von Regeln aus den trainierten Netzen geschehen, wodurch die Ansätze von modell- und datengetriebenen Analysen wieder näher aneinanderrücken, oder durch Verfahren zur Visualisierung der Beziehung von Eingangsdaten und Entscheidungen. Während dieser Ansatz für Bilddaten schon verbreitet ist, sind Lösungen für abstrakte Daten noch Forschungsgegenstand.

Handlungsempfehlungen

Die Interpretierbarkeit sollte von zwei Seiten betrachtet und auch angegangen werden. Zum einen sollten die Verfahren zur Entscheidungsfindung nachvollziehbarer werden, zum anderen sollte aber auch bei den verantwortlichen Entscheidern, die auf die Ergebnisse zugreifen, ein Verständnis über die Vorgehensweise von ML erarbeitet werden. Dementsprechend gilt es, Konzepte wie Transparenz und Erklärbarkeit von ML zu fördern und voranzutreiben. Zum anderen sollten aber auch Weiterbildungsangebote entwickelt werden, die Entscheidern die Grundlagen von ML vermitteln und sie in die Lage versetzt, Ergebnisse kritisch zu hinterfragen.

Daraus leiten sich die folgenden Handlungsempfehlungen an unterschiedliche Kreise ab:

- Politik:
  - Schaffen von Rahmenbedingungen, die interpretierbare ML begünstigen und beispielsweise fördern, Ergebnisse von ML durch Fachleute belastbar interpretierbar zu gestalten.

- Behörden:
  - Schaffen von Anforderungskatalogen zur Interpretierbarkeit von ML-Ergebnissen unter Berücksichtigung des Kenntnisstandes von erwarteten Endanwendern, aber auch Entwicklung von Konzepten zur Schulung von Anwendern im Umgang mit ML.
- Unternehmen:
  - Aus Anwendersicht entsprechen die Empfehlungen denen der Behörden. Aus Entwicklersicht gilt es, bei der Auswahl von ML solche Verfahren zu bevorzugen, die eine Interpretierbarkeit unterstützen und geeignete Schnittstellen und Visualisierungen zur Interpretation schaffen.
- Wissenschaft:
  - Entwickeln von sowohl ML-Verfahren, die einfacher zu interpretieren sind als beispielsweise die hochkomplexen Deep-Learning-Netze als auch von interpretierbaren Ergebnisdarstellungen, die auch abstrakte Daten jenseits multimedialer Inhalte umfassen.

6 Nachvollziehbarkeit des Einsatzes von KI

Die Nutzung von KI/ML wird von vielen Seiten mit Skepsis betrachtet, insbesondere auch deshalb, weil perspektivisch ihr autonomer Einsatz befürchtet wird. So entstehen sowohl ein Kontrollverlust bei einem Fehlverhalten der KI, aber auch ein Unvermögen durch das Objekt einer Entscheidung, zu erkennen, ob die Entscheidung von einem Menschen oder einer Maschine getroffen wurde. Noch stärker wird dieses Problem, wenn Ergebnisse der KI bewusst zur Täuschung eingesetzt werden. Deep Learning hat sogenannte Deep Fakes ermöglicht, durch die Menschen täuschend echt in Videos agieren, ihre Handlungen tatsächlich aber nur synthetisiert sind. Auch das synthetische Erzeugen von Texten und Klängen gewinnt stetig an Qualität.

Dies mündet in der Frage, wie ein Einsatz von KI erkannt werden kann, welche Bedeutung diese Information hat und wie diese Erkennung oder Kennzeichnung durchsetzbar ist. Hier sind aktive oder passive Strategien denkbar: Aktive Ansätze würden eine Kennzeichnung fordern, müssen dann aber auch Maßnahmen beim Versäumen der Kennzeichnung definieren. Passive Ansätze sind forensische Methoden, die bei gegebenen Daten anhand ihrer Eigenschaften eine Erkennung des Einsatzes von KI anstreben. Für Deep Fakes sind entsprechende Methoden bekannt, weisen aber durchaus signifikante Fehlerraten auf. Abhängig vom Testumfeld reichen die Erkennungsraten von grob 65 % bis zu fast 99 %. Beispielsweise erreichte bei der bekannten Facebook Deepfake Challenge der Gewinner eine Erkennungsrate von 82 %[13].

Handlungsempfehlungen

Die Methoden des ML, die Deep Learning und ähnliche Methoden zur künstlichen Erzeugung realitätsnaher Daten erlauben, führen in einigen Bereichen des öffentlichen Lebens bereits heute zu Verunsicherung, beispielsweise bei Desinformations-Kampagnen. Gleichzeitig führt dies dazu, dass die Abstreitbarkeit von Aufnahmen immer leichter wird, da diese als Fälschung abgetan werden können (liar's dividend). Dementsprechend ist es von Bedeutung, Ansätze zu verfolgen, die den Einsatz von KI beim Entstehen von Daten nachvollziehbar machen, insbesondere von solchen, die fälschlicherweise als von Menschen erstellt betrachtet werden.

Um sicherzustellen, dass Methoden des ML nicht unbemerkt eingesetzt werden, beispielsweise um Nutzer zu täuschen oder auch um Desinformationen zu verbreiten, sind unterschiedliche Strategien denkbar, die sich potentiell ergänzen. Aus technischer Sicht sind Methoden zur Erkennung von Inhalten, die durch ML entstanden sind, an geeigneten Stellen einzusetzen. Der umfassende Einsatz entsprechender Methoden kann durch Vorgaben durch die Gesetzgebung vorangetrieben werden. Da allerdings, wie oben beschrieben, nicht absehbar ist, ob eine zulässige Erkennung möglich ist, sollte auch über rechtliche Rahmenbedingungen für die Kennzeichnung des Einsatzes von ML nachgedacht werden. Hiermit wird der unbemerkte Einsatz von ML zwar technisch nicht weiter behindert, die Konsequenzen bei einer Aufdeckung eines verdeckten Einsatzes wären aber größer, was wiederum eine abschreckende Wirkung hat. Parallel dazu sollte auch die Strategie verfolgt werden, die Nachvollziehbarkeit von Informationen zu stärken, die nicht verändert worden sind. So ist als Alternative zur Aufdeckung von Deep Fakes denkbar, einen Rahmen zu schaffen, in welchem Originalaufnahmen nachweisbar als solche gekennzeichnet werden.

Daraus leiten sich die folgenden Handlungsempfehlungen an unterschiedliche Kreise ab:

- Politik:
  - Um die Kennzeichnung des Einsatzes von ML verbreitet durchzusetzen, sind gesetzliche Grundlagen notwendig, die Regeln für Einsatz und Kennzeichnung von ML vorgeben.
  - Gesetzlichen Rahmen schaffen, in dem Mediendaten nachweisbar als authentisch gekennzeichnet werden können.
  - Allgemein ist es in Anbetracht der Bedeutung von ML und seiner öffentlichen Wahrnehmung empfehlenswert, eine breite Aufklärung und Diskussion der

Aspekte von ML und die Abgrenzung von starker KI zu initiieren, um hier die Akzeptanz zu stärken, gleichzeitig aber auch das Risikobewusstsein zu schärfen.

- Behörden:
  - Hier sollten Kompetenzen und Werkzeuge aufgebaut werden, die den Stand der Möglichkeiten von ML abschätzen, um realistische Risiken abzuleiten und seinen Einsatz nach Möglichkeit zu erkennen.
  - Insbesondere dort, wo der Missbrauch von ML direkte Auswirkungen auf die Bürger hat, beispielsweise durch auf Deep Fakes basierenden Desinformationen, sind bei den Behörden Konzepte zur Aufdeckung zu erstellen.
- Unternehmen:
  - Wo ML eingesetzt wird und das Risiko besteht, nicht zwischen von Menschen oder ML erstellten Inhalten unterscheiden zu können, sei es im Kundenkontakt oder bei medialen Inhalten, sollte ein einheitliches Konzept zur Kennzeichnung dieser Inhalte entwickelt werden.
- Wissenschaft:
  - Entwicklung von Verfahren zur Erkennung des Einsatzes von ML insbesondere in Themenfeldern, bei denen der unerkannte Einsatz eine schädliche Wirkung haben kann.
  - Gleichzeitig aber auch der Entwurf von Strategien, wie andererseits Quellen, die nicht durch ML entstanden sind, verlässlich gekennzeichnet werden können.

1 https://www.theverge.com/2019/2/6/18213453/gmail-tensorflow-machine-learning-spam-100-million
2 Gangavarapu, T., Jaidhar, C.D. & Chanduka, B. Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev 53, 5019–5081 (2020). https://doi.org/10.1007/s10462-020-09814-9
3 https://spd.group/machine-learning/credit-card-fraud-detection/
4 https://www.netguru.com/blog/fraud-detection-with-machine-learning-banking
5 https://www.malwarepatrol.net/ai-enabled-phishing/
6 https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/
7 https://www.ibm.com/de-de/products/cognitive-security-analytics
8 Imam, Niddal H., and Vassilios G. Vassilakis. „A Survey of Attacks Against Twitter Spam Detectors in an Adversarial Environment.“ Robotics 8.3 (2019): 50.
9 Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. „Badnets: Identifying vulnerabilities in the machine learning model supply chain.“ arXiv preprint arXiv:1708.06733 (2017).
10 https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1
11 https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms
12 Kevin Rothrock. Facial Recognition Service Becomes a Weapon Against Russian Porn Actresses. 22. Apr. 2016. URL: https://advox.globalvoices.org/2016/04/22/facial-recognitionservice-becomes-a-weapon-against-russian-pornactresses/
13 https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Wissenschaftliche Arbeitsgruppe Nationaler Cyber-Sicherheitsrat

Seit Oktober 2018 unterstützt die Wissenschaftliche Arbeitsgruppe den Nationalen Cyber-Sicherheitsrat. Sie berät aus Perspektive der Forschung zu Entwicklungen und Herausforderungen im Hinblick auf eine sichere, vertrauenswürdige und nachhaltige Digitalisierung.

Mitglieder der Wissenschaftlichen Arbeitsgruppe sind: Prof. Dr. Claudia Eckert, Dr. Timo Hauschild, Prof. Dr. Jörn Müller-Quade, Prof. Dr.-Ing. Christof Paar, Prof. Dr. Gabi Dreo Rodosek, Prof. Dr. Alexander Roßnagel, Prof. Dr. Michael Waidner (Hauptautor dieses Impulspapiers)