Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Sicherer, robuster und nachvollziehbarer Einsatz von KI

Probleme, Maßnahmen und Handlungsbedarfe

Bundesamt für Sicherheit in der Informationstechnik Postfach 20 03 63
53133 Bonn
Tel.: +49 22899 9582- 0
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

1 Einleitung

Methoden der *Künstlichen Intelligenz* (KI) zeigen in vielen Anwendungsbereichen, wie zum Beispiel der Objekterkennung auf Bildern, sehr gute Leistungen und werden zunehmend in Bereichen eingesetzt, die unser tägliches Leben beeinflussen. Zu den Anwendungsgebieten gehören auch solche mit potentiell *kritischen Auswirkungen*, wie z. B. das (teil)autonome Fahren, die Gesichtserkennung oder die Auswertung medizinischer Daten. Gleichzeitig bestehen derzeit viele *ungelöste Probleme* hinsichtlich eines *sicheren, robusten und nachvollziehbaren Einsatzes* von KI.

Dieses Dokument präsentiert *ausgewählte Probleme* sowie *Maßnahmen* für einen solchen Einsatz in Bezug auf sogenannte *konnektionistische KI-Methoden* und zeigt *Handlungsbedarfe* aus Sicht des BSI auf. Ethische und rechtliche Fragestellungen werden im Folgenden nicht behandelt.

2 Begriffsdefinition und fachliche Einführung

Angelehnt an die Definition[1] der *hochrangigen Expertengruppe für KI der Europäischen Kommission* versteht das BSI unter dem Begriff *Künstliche Intelligenz* die Technologie und wissenschaftliche Disziplin, die mehrere Ansätze und Techniken wie z. B. maschinelles Lernen, maschinelles Schließen und die Robotik umfassen. KI-Systeme sind Software- und Hardwaresysteme, die Künstliche Intelligenz nutzen, um in der physischen oder digitalen Welt „rational" zu handeln. Auf Grundlage von Wahrnehmung und Analyse ihrer Umgebung agieren sie mit einem gewissen Grad an Autonomie, um bestimmte Ziele zu erreichen.

Sogenannte *konnektionistische KI-Methoden* basieren auf Modellen aus vielen einfachen Funktionseinheiten, die stark untereinander verknüpft sind, ähnlich wie Neuronen im menschlichen Gehirn. Das bekannteste Beispiel sind sogenannte tiefe *neuronale Netze*, welche i. d. R. Millionen von Parametern besitzen. Diese stellen die Eigenschaften von und Verbindungen zwischen den Neuronen dar. Die Struktur eines solchen Netzes und die Parameterwerte definieren das *KI-Modell* und kodieren dessen mögliche Reaktionen auf Eingaben implizit. Die Werte werden hierbei nicht manuell ermittelt, sondern maschinell mittels *Optimierungsverfahren* anhand von Trainingsdaten. Der Erfolg eines solchen Verfahrens wird mittels *Metriken* anhand von Testdaten quantifiziert. Die verwendeten Metriken sowie die Qualität und Quantität der Trainings- und Testdaten bestimmen die Funktionsweise des Modells wesentlich. Die Modelle sind zudem oft sehr *sensitiv*, d. h. bereits geringe Änderungen an den Eingabedaten können deren Verhalten wesentlich beeinflussen. Dies führt zu schwerwiegenden Folgen: Funktionsweise und Ausgaben eines solchen Modells sind aufgrund der impliziten Kodierung der möglichen Reaktionen auf Eingaben oft äußerst schwierig nachzuvollziehen, d. h. *es mangelt an Transparenz und Erklärbarkeit*. Zusammen mit der Sensitivität bewirkt dies, dass die *Robustheit* des Modells gegenüber zufälligen und gezielten Störungen nur *schwer* und in begrenztem Maße nachweislich *verifizierbar* ist. Daher ergeben sich neuartige Angriffsmöglichkeiten, welche im nächsten Abschnitt beschrieben werden.

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

3 Neuartige Angriffe

Konnektionistische KI-Methoden sind anfällig gegenüber qualitativ neuartigen Angriffen, die wir im Folgenden *KI-spezifische Angriffe* nennen. Diese Angriffe sind z. T. bereits durch legitime Anfragen an das KI-Modell möglich. Die derzeit relevantesten Angriffe dieser Art sind:

- **Evasion/Adversarial Attacks**: Durch eine Manipulation von Eingabedaten verleiten Angreifer das KI-Modell im Betrieb zu vom Entwickler nicht vorgesehenen Ausgaben. Das Modell selbst wird hierbei nicht verändert. Bereits geringfügige Änderungen der Eingabedaten, die schwierig zu detektieren und selbst für Menschen nicht unmittelbar erkennbar sind oder von ihnen als irrelevant interpretiert werden, können bedeutsame Auswirkungen haben.
- **Data Poisoning Attacks**: Durch eine Manipulation der Trainingsdaten des KI-Modells erwirken Angreifer, dass dieses auf (bestimmte) Eingaben nicht wie vom Entwickler vorgesehen reagiert. Aufgrund der vielen Daten und der mangelnden Transparenz sind diese Angriffe meist schwer detektierbar.
- **Privacy-Attacks**[2]: Angreifer extrahieren Informationen hinsichtlich der Trainingsdaten aus dem Modell. *Model Inversion Attacks* extrahieren Trainingsdaten und *Membership Inference Attacks* stellen fest, ob ein Datum zum Training verwendet wurde.
- **Model Stealing Attacks**: Angreifer extrahieren die Funktionalität des Modells. Dabei werden Informationen über die Struktur des Modells, z. B. relevante Entscheidungsparameter, extrahiert oder die Funktionalität des angegriffenen Modells (näherungsweise) kopiert. Ziel ist der Diebstahl geistigen Eigentums oder die Vorbereitung anderer Angriffe.

[2] Die Begriffe werden in der Literatur nicht einheitlich verwendet. Je nach Betrachtungsweise werden Privacy Attacks als Untermenge von Model Stealing Attacks oder als eigenständige Kategorie angesehen.

4 Maßnahmen zur Erhöhung der IT-Sicherheit

Gegenmaßnahmen für die KI-spezifischen Angriffe werden aktiv erforscht. Die derzeit bekannten Maßnahmen bieten jedoch großteils nur einen beschränkten Schutz: Beispielsweise können nach aktuellem Stand der Forschung alle bekannten Maßnahmen gegen Adversarial Attacks durch sogenannte *adaptive Angriffe*[3] überlistet werden. Dennoch können die existierenden Maßnahmen nützlich sein, um die Ausführung von Angriffen zumindest zu erschweren oder deren Auswirkungen zu mindern. Ob diese Verbesserungen ausreichen, muss im Kontext des jeweiligen Anwendungsfalls bewertet werden. Das Gefährdungspotenzial und damit die Notwendigkeit von Schutzmaßnahmen weisen dabei eine große Streuung auf, da sich sowohl die Auswirkungen von Fehlfunktionen und erfolgreichen Angriffen als auch die Randbedingungen des Einsatzes stark unterscheiden.

Professionelle Hersteller, Anbieter und Entwickler von konnektionistischen KI-Systemen sollten derzeit die folgenden Punkte berücksichtigen, um ein *Mindestlevel* an Sicherheit für die Systeme zu gewährleisten:

- Die *klassischen Maßnahmen hinsichtlich Software- und Systemsicherheit* gelten unverändert für KI-Systeme und sollten umgesetzt werden.[4] Dies alleine ist jedoch *nicht* ausreichend.
- Im Rahmen eines *KI-spezifischen Risikomanagements* sollte der *gesamte Lebenszyklus des KI-Systems* systematisch hinsichtlich relevanter Risiken analysiert werden, wobei die oben genannten KI-spezifischen Angriffe berücksichtigt werden sollten. Basierend auf der Analyse können risikobasiert Mitigationsmaßnahmen auf der Ebene des KI-Systems sowie weitere technische oder organisatorische Maßnahmen zur Änderung der Rahmenbedingungen abgeleitet werden. Die Robustheit gegenüber Adversarial Attacks kann z. B. im Rahmen eines sogenannten *adversarialen Trainings* verbessert werden. Die Wirksamkeit der Maßnahmen sollte auf Angemessenheit hinsichtlich des Anwendungsszenarios bewertet werden. Um Risiken valide bewerten zu können, kann es förderlich sein, selbst *adaptive Angriffe* auf die KI-Systeme *durchzuführen* (Red-Teaming) oder externe Parteien hierfür zu beauftragen. Die Risikoanalyse sollte *regelmäßig wiederholt* werden, um den aktuellen Stand der Forschung zu berücksichtigen.
- Die *Metriken*, welche verwendet werden, um die Qualität der KI-Modelle zu bewerten, sollten dem *Gefährdungspotenzial* der jeweiligen Anwendung *Rechnung tragen*. Neben der Genauigkeit auf den erwarteten Eingabedaten sollten auch andere Aspekte berücksichtigt werden, wie z. B. *Over-/Underfitting*[5], *Bias-Effekte*[6] sowie die *Robustheit gegenüber zufälligen oder gezielten Änderungen*. Im Idealfall werden verschiedene Modellansätze mittels unterschiedlicher Metriken verglichen und hinsichtlich ihrer Eignung für die jeweilige Anwendung bewertet.
- Eine *geeignete Qualität und Quantität der Trainings- und Testdaten* und ggf. notwendiger Betriebsdaten sollten durch systematische Untersuchungen und Maßnahmen sichergestellt werden. Es empfiehlt sich ein *professionelles Datenmanagement* einzuführen. Hierzu gehört z. B., dass Daten und Modelle gegen Manipulationen geschützt und Änderungen protokolliert werden sowie jedes Datum seiner Quelle zuzuordnen ist. Besondere Vorsicht ist bei der Verwendung von Daten oder Modellen aus externen Quellen geboten. Eine Entscheidung diesbezüglich muss das anwendungsspezifische Risiko berücksichtigen.

[3] Dies sind Angriffe, die spezifisch auf das jeweilige Modell und die genutzten Verteidigungsmaßnahmen angepasst werden.

[4] Konkrete Empfehlungen können beispielsweise dem IT-Grundschutz-Kompendium des BSI entnommen werden.

[5] Overfitting bezeichnet eine übermäßige und Underfitting eine unzureichende Anpassung eines KI-Modells an die Trainingsdaten. Beides wirkt sich negativ auf die Qualität des Modells aus.

[6] Diese entstehen durch systematische Verzerrungen innerhalb der verwendeten Trainingsdaten, in denen z. B. bestimmte Korrelationen häufiger auftreten, als dies in Wahrheit der Fall oder als es gesellschaftlich gewünscht ist.

- Um mögliche KI-spezifische *Angriffe zu detektieren* und Sicherheitsvorfälle nachvollziehen zu können, sollten Anfragen an und Zugriffe auf das KI-System geeignet *protokolliert* und die Protokolldaten regelmäßig auf Anomalien untersucht werden. Es sollten Prozesse etabliert werden, um auf Sicherheitsvorfälle im Betrieb zeitnah reagieren zu können.
- KI-Systeme sollten in *regelmäßigen Abständen* anhand der entsprechenden Metriken hinsichtlich einer korrekten Funktionsweise überprüft werden, um auf die *Veränderung von Umgebungsparametern* reagieren zu können.
- Die Kritikalität der *mangelnden Transparenz und Erklärbarkeit* von konnektionistischen Modellen sollte vor dem Hintergrund des jeweiligen Anwendungsfalls bewertet werden. Die Nutzung von sogenannten *XAI-Methoden*[7] kann unter Berücksichtigung ihrer aktuellen Limitationen in Erwägung gezogen werden, um die Ergebnisse in begrenztem Maße erklären zu können. Aus Sicht der IT-Sicherheit sind einfache und transparente Modelle gegenüber großen und komplexen vorzuziehen. Es empfiehlt sich zu prüfen, ob die Parameteranzahl reduziert werden kann oder intrinsisch interpretierbare KI-Modelle (z. B. Entscheidungsbäume), evtl. auch in Kombination, verwendet werden können.

Anbieter von KI-Systemen sollten zudem *präzise* und *verständlich* beschreiben, unter welchen *Randbedingungen* das KI-System welche *Funktionalität* aufweist und welche *Limitationen* das System hat. Diese Beschreibung sollte potentiellen Anwendern zugänglich gemacht werden, damit diese den Einsatz des KI-Systems für ihren Anwendungsfall bewerten können.

[7] Dies sind Methoden zur Interpretation von komplexeren KI-Modellen.

5 Handlungsbedarfe und Aktivitäten des BSI

Aus Sicht des BSI besteht derzeit ein dringender Handlungsbedarf, die Sicherheit von KI-Systemen weiter zu erforschen, um verlässliche Sicherheitsaussagen über die Systeme treffen zu können:

1. **Entwicklung von Standards, technischen Richtlinien, Prüfkriterien und Prüfmethoden**: Derzeit existieren keine hinreichend geeigneten Standards, um die Sicherheit von KI-Systemen für kritische Anwendungskontexte (wie sie z. B. in der Automobil- und Rüstungsindustrie, in der Biometrie, im Gesundheitswesen sowie im Finanz-, IT- und Telekommunikationsbereich vorliegen können) verlässlich zu bewerten und technisch zu prüfen. Auch für weniger kritische Anwendungen fehlen (mit wenigen Ausnahmen) Maßstäbe für die Sicherheit.
2. **Erforschung von wirksamen Gegenmaßnahmen gegen KI-spezifische Angriffe**: Die existierenden Maßnahmen für die oben genannten Angriffe sind oft nicht ausreichend. Um einen sicheren und robusten Betrieb von KI-Systemen zu ermöglichen, müssen weitere Gegenmaßnahmen möglichst praxisnah erforscht werden.
3. **Erforschung von Methoden der Transparenz und Erklärbarkeit**: Die oft mangelhafte Erklärbarkeit von KI-Systemen beeinflusst deren IT-Sicherheit maßgeblich und sorgt für fehlende Akzeptanz der Systeme seitens der Anwender. Es ist daher wichtig, auch die Methoden zur Erklärbarkeit praxisnah weiter zu erforschen.

Um die Sicherheit von KI-Systemen zu stärken, beteiligt sich das BSI an nationalen und internationalen Gremien und Gruppen. Das BSI arbeitet außerdem aktiv an der Entwicklung von Prüfkriterien und Prüfmethoden für KI-Anwendungen in unterschiedlichen Domänen, insbesondere im Bereich Automotive und Cloud-Services.

Eine tiefergehende Betrachtung ausgewählter Aspekte dieses Dokuments hat das BSI im Juli 2020 in einem Fachartikel[8] vorgenommen. Im Februar 2021 hat das BSI außerdem den AI Cloud Service Compliance Criteria Catalogue (AIC4)[9] veröffentlicht. Die AIC4-Kriterien definieren ein Basisniveau an Sicherheit für KI-basierte Cloud-Dienste und sind im Rahmen einer standardisierten Prüfung auditierbar. Ein entsprechender Prüfbericht kann, bei sachgemäßer Erstellung und Verwendung, professionelle Cloud-Kunden dabei unterstützen, die Informationssicherheit eines KI-Dienstes für die eigenen Anwendungsfälle zu bewerten.

Details zu den Aktivitäten und Publikationen des BSI können auf der Homepage[10] eingesehen werden.

[8] https://doi.org/10.3389/fdata.2020.00023
[9] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf
[10] https://www.bsi.bund.de/ki