Bedrohungen und Schutzmaßnahmen

ANGRIFF AUF DIE KÜNSTLICHE INTELLIGENZ

Künstliche Intelligenz ermöglicht es, komplexe Zusammenhänge und Muster aus großen Datenmengen zu extrahieren und in einem statistischen Modell zu erfassen. Anschließend lassen sich Aussagen über zukünftig auftretende Daten treffen. Mit dem zunehmenden Einsatz von künstlicher Intelligenz rücken solche Systeme auch immer mehr ins Visier von Cyberkriminellen. Unser Artikel beschreibt umfassend Angriffsszenarien und mögliche Abwehrmaßnahmen.

Die in den letzten Jahrzehnten gestiegene Rechenleistung hat die praktische Anwendung der künstlichen Intelligenz (KI) begünstigt. Auch erkennt man zunehmend das Potenzial der Technologie. So nutzen heute autonome Fahrzeuge eine KI, um Verkehrsschilder zu erfassen, Sprachassistenten verwenden sie für die Interpretation natürlicher Sprache und die Cybersicherheitsindustrie setzt sie für die Detektion von Cyberangriffen ein. Auch die zuletzt aufgekommene Diskussion um den Chatbot ChatGPT zeigt die Leistungsfähigkeit und zudem die daraus resultierenden gesellschaftlichen Folgen der KI.

Für das Training eines KI-Modells werden Trainingsdaten benötigt. Beim überwachten Lernen (engl. supervised learning) ist zusätzlich jedem Trainingsbeispiel ein Label zugeordnet, das der KI das gewünschte beziehungsweise erwartete Ergebnis zeigt. Basierend auf diesem Datensatz versucht die KI allgemeine Muster zu erfassen, mit denen sie auch neue Daten dem korrekten Label zuordnen kann. Mit neuen Daten sind hierbei Eingaben gemeint, die nicht in den Trainingsdaten enthalten sind. Zum Beispiel besteht der Trainingsdatensatz für einen Spamfilter aus E-Mails und jeder E-Mail könnte entweder ein Label „spam" oder „normal" zugeordnet werden. Der trainierte Spamfilter wird dann eingesetzt, um neue E-Mails zu klassifizieren.

Das zugrunde liegende Konzept der Daten, das gelernt werden soll, ist in der Regel jedoch unbekannt. In der Praxis steht nur eine empiri-

sche Verteilung (Stichprobe) der Grundwahrheit zur Verfügung. Die Herausforderung besteht nun darin, einen Trainingsdatensatz zusammenzustellen, der repräsentativ für die Grundwahrheit ist. Bei vielen Anwendungen sind nicht alle Prädiktoren messbar, sodass ein KI-Modell im Allgemeinen einen nicht reduzierbaren Fehler enthält [1]. Jedoch müssen KI-Modelle nicht perfekt sein, um nützlich zu sein. Entscheidend ist ein angemessener Kompromiss zwischen Nutzen, Kosten und Risiko.

Mit dem Einsatz von KI gehen aber auch Gefahren einher: Die Konsequenzen einer Fehlentscheidung können sich je nach Anwendungsfall von finanziellen Schäden bis hin zu tödlichen Folgen erstrecken. Gleichzeitig ist es ein offenes Problem, die korrekte Funktionsweise eines KI-Modells sowie die zuverlässige Erkennung von Fehlern nachzuweisen. Eine KI ist meistens fehlerbehaftet, sodass es immer ein Restrisiko von Fehlentscheidungen gibt, mit dem man umgehen muss. Das Design eines guten KI-Modells erfordert daher zahlreiche wohlüberlegte Entscheidungen und Prozesse.

Wie die meisten IT-Systeme weisen auch KI-Modelle konzeptionelle Schwachstellen auf, die ein bösartiger Akteur ausnutzen kann, um ein KI-Modell zu manipulieren und somit Entscheidungen zu beeinflussen.

ANGRIFFE AUF KI-MODELLE

In den letzten Jahren haben sich zahlreiche Studien mit den Schwachstellen von KI-Modellen befasst und mögliche Angriffe gezeigt. Auch sind bereits viele Vorfälle im Zusammenhang mit KI-Technologie dokumentiert [2].

Unter Laborbedingungen entwickelte und evaluierte KI-Modelle weisen bei der produktiven Anwendung oft Schwächen auf. So können in der Praxis Daten von einem Angreifer stammen – es ist daher unrealistisch anzunehmen, dass Daten von externen Quellen vertrauenswürdig sind. Ein realistisches Angriffsszenario sollte die Existenz eines Angreifers annehmen, der Daten manipulieren und mit einem KI-Modell interagieren kann. In diesem Modell sind folgende Bedrohungen zu berücksichtigen:

- Durch Veränderung der Trainingsdaten kann ein Angreifer Einfluss auf den Entscheidungsprozess eines KI-Modells nehmen und so Fehlentscheidungen verursachen.
- Durch Veränderung einer Eingabe (unter Nutzung des gesamten Eingaberaums) lässt sich die Position im Eigenschaftsraum verschieben, sodass ein KI-Modell diese Eingabe falsch klassifiziert.
- Bei der Interaktion mit einem KI-Modell kann ein Angreifer Eingabe-Ausgabe-Paare sammeln, die Aufschluss über die Funktionsweise und die verwendeten Trainingsdaten geben können.

In Abbildung 1 sind der Aufbau einer KI-Pipeline und potenziell angreifbare Komponenten dargestellt. Im Folgenden stellen wir die gängigsten Angriffsvektoren auf KI-Modelle vor, die bei nahezu allen KI-basierten Anwendungen berücksichtigt werden sollten.

POISONING-ANGRIFF

Bei einem Poisoning-Angriff manipuliert ein Angreifer die Trainingsdaten, um die Leistung einer KI zu verschlechtern. Schon die Manipulation weniger Daten kann einen weitreichenden Einfluss auf eine KI haben [3]. Zum Beispiel könnte ein Angreifer die Labels von Verkehrsschildern ändern, sodass das KI-Modell häufiger falsche Klassifizierungen ausgibt.

Die Voraussetzung für einen Poisoning-Angriff ist ein direkter oder indirekter Zugriff auf die Daten, die ein KI-Modell für das Training verwendet. Das lässt sich auf folgende Arten bewerkstelligen:

- Manipulation eines Trainingsdatensatzes bei der Übertragung über einen unsicheren Kommunikationskanal und fehlender Integritätsüberprüfung auf der Empfängerseite.
- Erstellung eines Trainingsdatensatzes mit falschen Labels, der anschließend öffentlich zur Verfügung gestellt wird. Hierbei wird darauf spekuliert, dass automatisierte Systeme diese Daten sammeln und sie ohne Überprüfung an ein KI-Modell zum Trainieren weitergeleitet werden; oder dass beim manuellen Bezug der Daten durch einen Menschen die fehlerhaften Labels nicht auffallen.
- Manipulation von Daten und/oder Labels eines existierenden Datensatzes unter Ausnutzung von unzureichender Zugriffskontrolle und Authentifizierung oder Sicherheitslücken in Software.
- Kompromittierung eines Datenanbieters oder Label-Erstellers.
- Unterwanderung eines Crowdsourcing-Anbieters.
- Im Fall von KI-Modellen zur Detektion von Cyberangriffen hat ein Angreifer (teilweise) Kontrolle über die Trainingsdaten, weil der Angreifer die Entität darstellt, deren Verhalten gelernt werden soll.

SCHUTZ VOR POISONING-ANGRIFFEN

Um eine KI vor Poisoning-Angriffen zu schützen, müssen die Trainingsdaten vor unautorisierter Veränderung gesichert werden. Da diese Daten

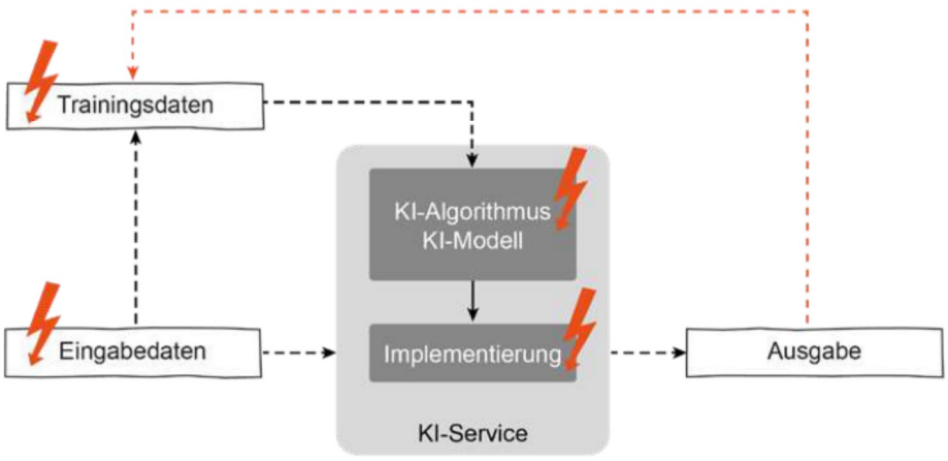

Abbildung 1: Beispielhafte KI-Pipeline mit potenziellen Angriffsflächen

häufig aus externen Quellen bezogen werden, sollte zusätzlich eine Beurteilung der Qualität und der Vertrauenswürdigkeit der Trainingsdaten erfolgen. Die folgenden Punkte beschreiben mögliche Schutzmaßnahmen:

- Die Trainingsdaten sollten über einen sicheren Kommunikationskanal übertragen werden. Zusätzlich sollte die Integrität eines Trainingsdatensatzes überprüft werden.

- Trainingsdaten sollten nicht wahllos gesammelt werden, sondern von vertrauenswürdigen Datenquellen bezogen werden. Außerdem sollte auch überprüft werden, ob die Daten geeignet und repräsentativ für die zu lernende Aufgabe sind. Idealerweise steht ein Datasheet [4] zur Verfügung, welches den Datensatz dokumentiert.

- Trainingsdaten sollten vor Manipulationen geschützt werden. Hierzu sollte sowohl ein Zugriffsmanagement als auch ein geeignetes Authentifizierungsverfahren implementiert werden. Auch sollten sicherheitskritische Software-Patches schnell angewendet werden. Eine Versionierung der Trainingsdaten ermöglicht es, nach einem Vorfall auf einen vertrauenswürdigen Datensatz zurückzugreifen. Eine Protokollierung von Änderungen der Trainingsdaten erleichtert die Untersuchung eines Vorfalls.

- Beim Bezug von Daten oder Labels aus externen Quellen sowie bei der Verwendung und Einbindung von externen Ressourcen in die KI-Pipeline sollten Risiken in der KI-Lieferkette identifiziert und eingeschätzt werden.

- Bevor ein KI-Modell produktiv eingesetzt wird, sollte dessen Leistung auf einem Testdatensatz (Evaluationsdatensatz) getestet werden. Dieser Testdatensatz sollte die Grundwahrheit möglichst gut repräsentieren und darf nicht im Training verwendet worden sein. Für die Modell-Evaluation sind zum Anwendungsfall passende Metriken auszuwählen und zu messen. Diese Metriken sollten auch kontinuierlich während des Modelleinsatzes beobachtet werden, um einen Leistungsabfall des KI-Modells erkennen zu können.

Ein Leistungsabfall kann ein Indikator für einen Poisoning-Angriff sein und einen Prozess zur Untersuchung des Trainingsdatensatzes sowie der KI-Lieferkette anstoßen.

EVASION-ANGRIFF

Ziel eines Evasion-Angriffs ist die Erstellung einer Eingabe, die eine falsche oder spezifische Entscheidung verursacht. Eine solche Eingabe wird als „Adversarial Example“ [5] bezeichnet. Zum Beispiel könnte ein Angreifer eine spezielle Brille herstellen und tragen, um sich gegenüber einem Gesichtserkennungssystem als eine andere Person auszugeben [6].

Voraussetzungen zur Durchführung eines solchen Angriffs sind grundlegende Informationen über und (in)direkter Zugriff auf ein KI-Modell. In der Regel sind einem Angreifer grobe Informationen über die Aufgabe eines KI-Modells, die Repräsentation von Eigenschaften und die Art der Trainingsdaten bekannt [7]. Der Zugriff auf ein KI-Modell kann je nach Anwendungsszenario unterschiedlich gestaltet sein:

- Wenn ein KI-Modell öffentlich verfügbar ist, von einem Modell-Anbieter an mehrere Kunden verkauft wird oder mit einem Cyberangriff gestohlen werden kann, hat ein Angreifer direkten Zugriff auf das Ziel-Modell und kann Adversarial Examples lokal testen.

- Es steht ein Orakel (vgl. Infobox) zur Verfügung, sodass ein Angreifer Anfragen an die KI senden kann. In diesem Fall hat er indirekten Zugriff auf das Ziel-Modell und muss zum Testen von Adversarial Examples das Orakel nutzen. Die Antworten des Orakels kann ein Angreifer sammeln und beliebig verwenden, zum Beispiel um seine Strategie anzupassen.

- Der Angreifer hat weder direkten noch indirekten Zugriff auf das Ziel-Modell und nur einige sehr wenige Versuche, eine Fehlentscheidung zu verursachen, zum Beispiel beim Entsperren eines gestohlenen Smartphones mit einem Adversarial Example (Gesichtserkennung).

Zusätzlich besteht oftmals die Möglichkeit, mit öffentlichen Datensätzen ein ähnliches KI-Modell zu trainieren und dieses zur Vorbereitung eines Adversarial Example zu verwenden. Zu diesem Zweck wird die sogenannte Übertragbarkeitseigenschaft (engl. transferability property) verwendet [8]. In der KI-Domäne besagt diese Eigenschaft, dass ein Adversarial Example, das bei einem KI-Modell funktioniert, mit hoher Wahrscheinlichkeit auch bei einem anderen KI-Modell funktionieren wird, wenn beide KI-Modelle auf die gleiche Aufgabe trainiert wurden.

ORAKEL

Mit dem Begriff Orakel ist ein System gemeint, dass es ermöglicht Informationen herauszufinden, welche eigentlich nicht direkt zugänglich sind. Zum Beispiel wird angenommen, dass die interne Funktionsweise eines Spamfilters für einen Angreifer unbekannt ist. Ansonsten könnte der Spamfilter trivial umgangen werden. In der Praxis kann ein Angreifer jedoch E-Mails versenden (das Orakel befragen) und sehen, welche von diesen E-Mails im Spamordner landen (das Orakel antwortet). Das ermöglicht einem Angreifer abzuleiten, welchen Einfluss bestimmte Wörter auf die Entscheidung des Spamfilters haben.

Effektiv ist das mit einem direkten Zugriff auf das Ziel-Modell gleichzusetzen.

SCHUTZ VOR EVASION-ANGRIFFEN

Die Annahme, dass durch Geheimhaltung von Details über die Implementierung und des Trainings eines KI-Modells Angriffe erheblich erschwert werden, ist fraglich. Denn schon durch Kenntnis der Aufgabe eines KI-Modells, lassen sich Informationen wie Modell-Architektur, Trainingsalgorithmus, Art der Trainingsdaten und die Eigenschaftsrepräsentation (zumindest ungefähr) ableiten. Des Weiteren steht für die meisten Anwendungsszenarien ein entsprechender Trainingsdatensatz öffentlich zur Verfügung. In der Praxis hat ein Angreifer oftmals auch Zugriff auf eine Ein-Ausgabe-Schnittstelle. Sinnvolle Schutzmaßnahmen sollten dieses Angriffsmodell berücksichtigen und diesem standhalten.

Eine Verwundbarkeit gegen Adversarial Examples lässt sich nie ganz ausschließen. Wie bereits erwähnt, sind KI-Modelle in der Regel fehlerbehaftet, weil nicht alle Prädiktoren erfassbar sind. Es stellt eine Herausforderung dar, während des Trainings alle Areale im Eigenschaftsraum zu explorieren. Zudem gibt es in vielen Domänen Grenzfälle, für die eine eindeutige Vorhersage schwierig ist. Manchmal sind sich selbst Domänenexperten nicht einig, welches Label zu vergeben ist. Besonders wenn konkurrierendes Verhalten, zum Beispiel die Erkennung von Spam, gelernt werden soll, ist anzunehmen, dass Adversarial Examples nicht ausgeschlossen werden können.

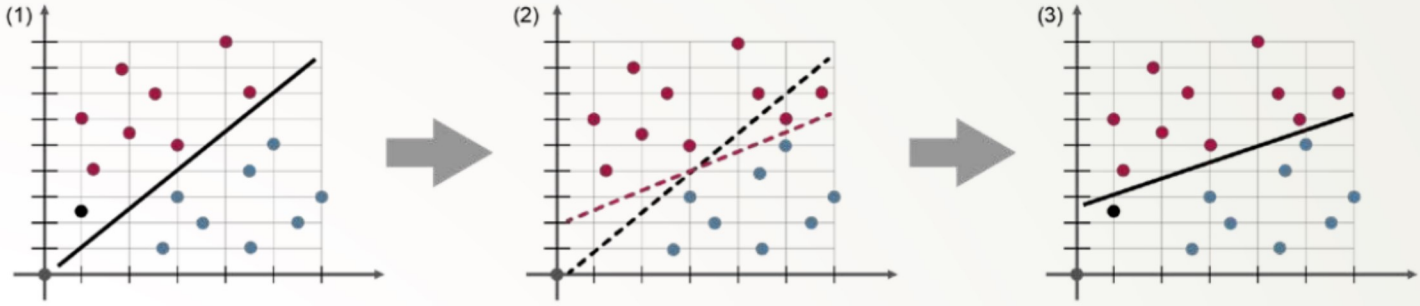

Abbildung 2: Die Manipulation einer Entscheidungsgrenze verändert die Entscheidung für eine bestimmte Eingabe.

Daher ist ein gutes Verständnis der vorliegenden Daten und des zu lösenden Problems unerlässlich für das Design eines robusten KI-Modells. Aufgrund der Vielzahl an Einsatzmöglichkeiten für KI, ist die Erstellung einer universellen Liste von Schutzmaßnahmen nicht möglich. Jedoch gibt es einige Maßnahmen, die sich auf viele KI-Modelle anwenden lassen:

- Zusätzlich zu einer quantitativen Auswertung mittels Metriken sollten auch Methoden der erklärbaren KI (engl. explainable AI) eingesetzt werden, um den Entscheidungsprozess eines KI-Modells nachvollziehen zu können. Hierdurch kann erkannt werden, wenn ein KI-Modell irrelevante Korrelationen als Prädiktoren verwendet, die keiner Kausalität in der jeweiligen Domäne entsprechen. Wenn solche Korrelationen gelernt werden, deutet das darauf hin, dass die Trainingsdaten die Grundwahrheit nicht ausreichend repräsentieren. Zum Beispiel, wenn ein Husky auf einem Bild aufgrund von Schnee im Hintergrund als Wolf klassifiziert wird, könnte dies daran liegen, dass in den Trainingsdaten keine Bilder von Wölfen ohne Schnee im Hintergrund vorliegen [9].

- Eingaben und zugehörige Vorhersagen im produktiven Betrieb sollten (stichprobenartig) gesammelt werden, um unerwartetes Verhalten oder Modellalterung erkennen zu können. Mit diesen Informationen können Schwachstellen aufgedeckt werden oder sich die Notwendigkeit einer Aktualisierung der Trainingsdaten ergeben.

- Alternativ oder ergänzend kann dem Nutzer eines KI-Modells eine Funktion zur Verfügung gestellt werden, mit der falsche Vorhersagen gemeldet werden können. Alle problematischen Eingaben können in Zukunft in die Evaluation und Qualitätssicherung einfließen.

- Software-Bibliotheken, die Algorithmen zur Generierung von Adversarial Examples implementieren, können genutzt werden, um die Robustheit eines KI-Modells zu testen.

- Je nach dem Grad der Autonomie und der Kritikalität einer KI-basierten Anwendung kann es sinnvoll sein, dass weiterhin ein Mensch die Kontrolle über die finale Entscheidung behält. Dabei sind die Nachvollziehbarkeit eines Ergebnisses sowie die Kommunikation von Konfidenzwerten (quantitative Sicherheit einer Empfehlung oder Entscheidung) von wichtiger Bedeutung, um dem Benutzer eine informierte Entscheidung zu ermöglichen. Jedoch sollte auch die Gefahr eines „Automation Bias“ nicht außer Acht gelassen werden.

Ein vorangegangener Poisoning-Angriff kann einen Evasion-Angriff begünstigen: Indem die Gewichtung von Prädiktoren verändert wird, kann die Erstellung von Adversarial Examples einfacher ausfallen oder bestimmte Eingaben können zu einem Adversarial Example werden. Zum Beispiel kann ein Angreifer gezielte E-Mails senden, sodass sich ein Spamfilter so verändert, dass eine bestimmte Spammail übersehen wird, die zuvor erkannt worden wäre [10]. In Abbildung 2 ist dieses Vorgehen konzeptionell dargestellt. Eine spezielle Ausprägung der Kombination dieser beiden Angriffsvektoren ist ein Backdoor-Angriff.

BACKDOOR-ANGRIFF

Bei einem Backdoor-Angriff wird zunächst ein Poisoning-Angriff durchgeführt, um einen Evasion-Angriff vorzubereiten. Die Trainingsdaten werden hierbei so manipuliert, dass eine gezielte Fehlentscheidung verursacht werden kann, wenn ein bestimmter Auslöser (engl. backdoor trigger [11]) in einer Eingabe vorhanden ist. In Abwesenheit des Auslösers verhält sich das KI-Modell aber unverändert.

Zum Beispiel könnte ein Angreifer Bilder von verschiedenen Verkehrsschildern mit einem Sticker und dem Label „Vorfahrt“ in einen Trainingsdatensatz einfügen. Dadurch wird das KI-Modell lernen, den Sticker mit dem Label „Vorfahrt“ zu assoziieren. Als Konsequenz kann durch Einfügen des Stickers jedes Bild zu einem Adversarial Example werden, beispielsweise ein Stoppschild (vgl. Abbildung 3).

Abbildung 3: Stoppschild mit einem Auslöser, der zu einer falschen Klassifizierung als Vorfahrtsschild führt.

Dieser Angriffsvektor unterstreicht noch einmal die Kritikalität der Schutzmaßnahmen vor Poisoning-Angriffen. Zusätzlich sollte man beachten, dass vortrainierte Modelle von externen Quellen möglicherweise mit manipulierten Daten trainiert wurden oder sogar bereits eine Backdoor enthalten können.

SCHUTZ VOR BACKDOOR-ANGRIFFEN

Bei der Verwendung von vortrainierten Modellen von externen Quellen sind folgende Maßnahmen zu empfehlen:

- Ein vortrainiertes Modell sollte ausschließlich über einen sicheren Kommunikationskanal übertragen und nach Empfang sollte dessen Integrität überprüft werden.

- Ein vortrainiertes Modell sollte nach Möglichkeit von einer vertrauenswürdigen Quelle bezogen werden. Idealerweise ist eine Model Card [12] vorhanden, die das Modell dokumentiert – und die verwendeten Trainingsdaten sind einsehbar oder zumindest dokumentiert.

Des Weiteren können die in den vorherigen Kapiteln beschriebenen Schutzmaßnahmen für einen Backdoor-Angriff übernommen werden, da dieser Angriffsvektor eine Kombination eines Poisoning-Angriffs und eines Evasion-Angriffs ist.

MODEL-EXTRACTION-ANGRIFF

Ziel eines Model-Extraction-Angriffs ist die Erstellung einer lokalen Kopie eines KI-Modells unter Verwendung eines Orakels. Zum Beispiel kann ein Unternehmen ein KI-Modell über eine Cloud-API anbieten. Auf diese Weise können Kunden die Funktion in ihre Anwendungen einbinden, ohne dass es lokal vorliegen muss. Das wertvolle geistige Eigentum verbleibt beim Anbieter-Unternehmen. Für einen Angreifer stellt dieses Szenario jedoch ein Orakel dar, womit sich das KI-Modell aus der Cloud kopieren lässt.

Voraussetzung zur Durchführung eines Model-Extraction-Angriffs ist ein indirekter Zugriff auf das Ziel-Modell über ein Orakel. Das grundlegende Vorgehen, zum Beispiel bei neuronalen Netzen, besteht darin, Eingaben zu finden, die Grenzfälle für das Ziel-Modell darstellen und somit einen Konfidenzwert nahe 0,5 verursachen. Weil sich diese Eingaben im Eigenschaftsraum sehr nahe an der Entscheidungsgrenze (engl. decision boundary) befinden, kann mit ihnen das Ziel-Modell beziehunsgweise eine Approximation dessen rekonstruiert werden. Mit anderen Methoden ist das auch bei Entscheidungsbäumen und linearer Regression möglich.[13]

Ein erfolgreich extrahiertes KI-Modell kann wiederum einen Evasion-Angriff begünstigen, da aufgrund der Übertragbarkeitseigenschaft eine Annäherung ausreicht. Weitaus schwerwiegender ist, dass durch so einen Angriff geistiges Eigentum gestohlen werden kann.

SCHUTZ VOR MODEL-EXTRACTION-ANGRIFFEN

Da ein Model-Extraction-Angriff auf einem Orakel-Zugriff basiert, beschränken sich die Maßnahmen auf die Konfiguration dieses Orakels. Auf technischer Ebene besteht die Ausgabe eines KI-Modells aus einer Vielzahl von Informationen. Man stelle sich ein KI-Modell zur Klassifizierung von Objekten auf Bildern vor. Die Ausgabe besteht hier aus einem Vektor von Konfidenz- beziehungsweise Wahrscheinlichkeitswerten. Das kann man sich als eine Liste von Label-Konfidenz-Tupel für alle Objekte vorstellen, auf die das Modell trainiert wurde und die es daher erkennen kann. In vielen Fällen ist jedoch nur ein Teil dieser Informationen für eine Anwendung oder einen Endnutzer relevant. Folglich besteht die Schutzmaßnahme vor einem Model-Extraction-Angriff darin, die vom Orakel ausgegebenen Informationen zu minimieren:

- Besonders wenn ein KI-Modell viele Labels gelernt hat, ist eine Minimierung der Anzahl der ausgegebenen Labels sinnvoll, weil oft nur die Labels, die über einem bestimmten Konfidenz-Schwellenwert liegen, von Interesse sind. Je nach Anwendungsfall kann alternativ immer eine statische Anzahl von Labels mit den höchsten Konfidenzwerten zurückgegeben werden, zum Beispiel die Top-10-Labels.

- Konfidenzwerte könnten gerundet oder sogar entfernt werden, sodass ausschließlich Labels ausgegeben werden. Dadurch wird es dem (legitimen) Benutzer aber auch erschwert, die (Un)Sicherheit einer Entscheidung einzuschätzen.

Auch wenn die Ausgaben eines Orakels minimal sind, kann ein Angreifer noch das Label erfahren, und ein Model-Extraction-Angriff ist weiterhin möglich.

RICHTLINIEN UND REGULIERUNG

Im Rahmen der europäischen Strategie für KI, hat die EU-Kommission eine unabhängige Expertengruppe damit beauftragt, Ethik-Leitlinien für eine vertrauenswürdige KI zu erarbeiten [14]. Die Ergebnisse wurden im April 2019 veröffentlicht und umfassen sieben Leitlinien [15]:

- Vorrang menschlichen Handelns und menschliche Aufsicht
- Technische Robustheit und Sicherheit
- Datenschutz und Datenqualitätsmanagement
- Transparenz
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen
- Rechenschaftspflicht

Ergänzend wurde eine Bewertungsliste erstellt, die bei der Umsetzung dieser Leitlinien unterstützt [16].

Im Februar 2020 wurde im „Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen" ein risikobasierter Ansatz zur Bewertung von KI-Anwendungen vorgeschlagen. Demnach soll eine Risikoeinstufung die regulatorischen Anforderungen bestimmen. Mit diesem Ansatz will man einen verhältnismäßigen Rechtsrahmen in der EU schaffen. Für KI-Anwendungen, die ein hohes Risiko darstellen, werden Anforderungen an die folgenden Bereiche vorgeschlagen:

- Trainingsdaten
- Aufbewahrung von Daten und Aufzeichnungen
- Vorzulegende Informationen
- Robustheit und Genauigkeit
- Menschliche Aufsicht
- Besondere Anforderungen an bestimmte KI-Anwendungen

Der Entwurf für ein „Gesetz über künstliche Intelligenz" von April 2021 greift diese Vorarbeiten auf und definiert drei Risikostufen [17]:

- KI-Anwendungen mit einem unannehmbaren Risiko sollen verboten werden.

- KI-Anwendungen mit einem hohen Risiko sollen obligatorischen Anforderungen unterliegen. Zum Beispiel soll für Hochrisiko-KI-Systeme eine Konformitätsbewertung verpflichtend sein, die eine Umsetzung definierter Anforderungen nachweist, zum Beispiel ein Risikomanagementsystem, technische Dokumentation, menschliche Aufsicht und Cybersicherheit. Des Weiteren sollen Anbieter verpflichtet werden, ihre Hochrisiko-KI-Systeme auch nach dem Inverkehrbringen zu beobachten und Vorfälle beziehungsweise Fehlfunktionen zu melden.

- Alle anderen KI-Anwendungen mit einem geringen oder minimalen Risiko unterliegen nur wenigen Transparenzpflichten oder sind von diesem Gesetz ausgenommen [18]. Bei KI-Systemen, die mit Nutzern interagieren

oder Inhalte erzeugen, besteht das Risiko, dass Menschen manipuliert werden könnten. Beispiele hierfür sind Chatbots, Emotionserkennungssysteme oder die Darstellung von künstlich erzeugten Bild-, Ton- oder Videoinhalten, die authentisch wirken. In solchen Fällen soll den Nutzern klar kommuniziert werden, dass es sich um ein KI-System handelt, unabhängig von der Risikostufe.

Ein im Dezember 2021 veröffentlichter Bericht der EU-Cybersicherheitsagentur (ENISA) gibt einen umfassenden Literaturüberblick über künstliche Intelligenz [19]. Insbesondere werden in dem Bericht Bedrohungen für KI-Anwendungen und Schutzmaßnahmen aufgeführt.

Zudem enthält die im Februar 2022 veröffentlichte Richtlinie VDE SPEC 90012 V1.0 eine Liste von Fragen, mit denen KI-Anwendungen in den Bereichen Transparency, Accountability, Privacy, Fairness und Reliability bewertet und verglichen werden können. Zum Beispiel wird abgefragt, wie detailliert die Eigenschaften der Daten und KI-Modelle dokumentiert sind und welche Maßnahmen eingerichtet sind, um die Integrität, Robustheit und Cybersicherheit des KI-Systems zu gewährleisten.

FAZIT

Das Verhalten eines KI-Modells wird von den Daten bestimmt, mit denen es trainiert wurde. Wichtig für ein vertrauenswürdiges KI-Modell sind daher die Qualität der Trainingsdaten und eine Evaluation, ob die zugrunde liegenden Konzepte einer Domäne im gelernten Entscheidungsprozess korrekt abgebildet werden. In der Regel sind KI-Modelle fehlerbehaftet, sodass immer die Möglichkeit von Fehlentscheidungen bei Grenzfällen besteht.

Wenn ein KI-Modell eingesetzt wird, sind neue Angriffsvektoren zu berücksichtigen. So kann man zum Beispiel über eine Manipulation der Trainings- oder Eingabedaten die Entscheidungen eines KI-Modells beeinflussen und über eine Interaktion mit einem KI-Modell kann ein Angreifer auf die interne Funktionsweise schließen und sogar eine Kopie erstellen.

Diese Bedrohungen sollten Entwickler schon beim Design eines KI-Modells berücksichtigen und potenzielle Risiken einschätzen. Die vorgestellten Schutzmaßnahmen können dabei helfen, die Angriffsfläche und somit die Risiken zu minimieren. ■

Literatur

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. Springer, 2021.

[2] https://incidentdatabase.ai/

[3] B. Biggio, B. Nelson, and P. Laskov, „Poisoning Attacks against Support Vector Machines", in International Conference on Machine Learning, ser. ICML '12. Edinburgh, Scotland, GB: Omnipress, Jun. 2012, pp. 1807–1814.

[4] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, „Datasheets for Datasets", Communications of the ACM, vol. 64, no. 12, pp. 86–92, Nov. 2021.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, „Intriguing Properties of Neural Networks", in International Conference on Learning Representations, ser. ICLR '14. Banff, Alberta, Canada: IEEE, Apr. 2014, pp. 372–387.

[6] 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 ACM Conference on Computer and Communications Security, ser. CCS '16. Vienna, Austria: ACM, Oct. 2016, pp. 1528–1540.

[7] B. Biggio and F. Roli, „Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning", Pattern Recognition, vol. 84, pp. 317–331, Dec. 2018.

[8] N. Papernot, P. McDaniel, and I. Goodfellow, „Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples", arXiv, vol. abs/1605.07277, pp. 1–13, May 2016.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, „"Why Should I Trust You?": Explaining the Predictions of Any Classifier", in ACM International Conference on Knowledge Discovery in Data Mining, ser. KDD '16. San Francisco, California, USA: ACM, Aug. 2016, pp. 1135–1144.

[10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, „The Security of Machine Learning", Machine Learning, vol. 81, no. 2, pp. 121–148, Nov. 2010.

[11] T. Gu, B. Dolan-Gavitt, and S. Garg, „BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", arXiv, vol. abs/1708.06733v2, pp. 1–13, Mar. 2019.

[12] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, „Model Cards for Model Reporting", in Conference on Fairness, Accountability, and Transparency, ser. FAT* '19. Atlanta, GA, USA: ACM, Jan. 2019, pp. 220–229.

[13] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, „Stealing Machine Learning Models via Prediction APIs", in USENIX Security Symposium, ser. SSYM '16. Austin, Texas, USA: USENIX, Aug. 2016, pp. 601–618.

[14] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=COM%3A2018%3A237%3AFIN

[15] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[16] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[17] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206

[18] https://ec.europa.eu/commission/presscorner/detail/de/ip_21_1682

[19] https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

DOMINIK ADLER
ist Wissenschaftlicher Mitarbeiter im Institut für Internet-Sicherheit – if(is) an der Westfälischen Hochschule in Gelsenkirchen und beschäftigt sich mit Angriffen auf künstliche Intelligenz.

NURULLAH DEMIR
ist Wissenschaftlicher Mitarbeiter im Institut für Internet-Sicherheit – if(is) an der Westfälischen Hochschule in Gelsenkirchen und beschäftigt sich mit Angriffen auf künstliche Intelligenz.

NORBERT POHLMANN
ist Professor für Cybersicherheit und Leiter des Instituts für Internet-Sicherheit – if(is) an der Westfälischen Hochschule in Gelsenkirchen sowie Vorstandsvorsitzender des Bundesverbands IT-Sicherheit – TeleTrusT und im Vorstand des Internetverbandes – eco.