A-SIT

Zentrum für sichere Informationstechnologie - Austria

Sichere Entwicklung von KI-Diensten

Sichere Entwicklung von KI-Diensten

Autor:
Bianca Danczul
Mail: bianca.danczul@iaik.tugraz.at
Datum: 15.02.2023

Abstract/Zusammenfassung:
KI-basierte Services können den Zeitaufwand und die Kosten auf Seiten der Behörden und privaten Unternehmen reduzieren, während die Zufriedenheit der Nutzer*innen erhöht wird. Sie bergen aber naturgemäß auch Risiken, vor allem, wenn persönlich identifizierbare Information (PII) verarbeitet werden. Dementsprechend ist es wichtig, dass bei deren Entwicklung auch Security und Compliance-Anforderungen betrachtet und eingehalten werden.

In der klassischen Softwareentwicklung gibt es zu diesem Zwecke Frameworks, die die sichere Entwicklung von Diensten vereinfachen. Diese sind jedoch nicht eins zu eins auf AI-basierte Dienste anwendbar. Ziel der vorliegenden Studie ist es daher, herauszufinden, welche Möglichkeiten für die sichere Entwicklung von AI-Diensten existieren und ob deren Umsetzung auch in der Praxis erfolgt. Die Ergebnisse werden aufgearbeitet und Raum für weitere Forschung sowie Lösungsvorschläge werden aufgezeigt.

Inhaltsverzeichnis

1. Einleitung

Künstliche Intelligenz (KI, engl. AI) ist dabei, die Welt, in der wir leben, zu revolutionieren und sowohl im privaten als auch im öffentlichen Sektor bedeutende Fortschritte gebracht. Aufgrund der vermehrten Anwendung von KI-Applikationen und angesichts der heutigen vernetzten Welt, in der Cyber-Bedrohungen und Datenschutzverletzungen immer häufiger vorkommen, kann die Wichtigkeit der sicheren Entwicklung von KI-Anwendungen nicht überbewertet werden.

Die Entwicklung sicherer KI-Anwendungen ist entscheidend, um die Integrität der Anwendung zu schützen, Datenschutzverletzungen zu verhindern und die Privatsphäre der Benutzer*innen zu wahren. Die potenziellen Auswirkungen einer Sicherheitsverletzung in einer KI-Anwendung können schwerwiegend sein und zu erheblichen finanziellen und Reputationsverlusten führen. Daher ist es wichtig, während des gesamten Entwicklungsprozesses Sicherheitsanforderungen zu betrachten, um die Sicherheit und Zuverlässigkeit von KI-Anwendungen zu gewährleisten und das Vertrauen der Benutzer*innen, die sich im Alltag auf diese Anwendungen verlassen, zu stärken.

Bei der herkömmlichen Softwareentwicklung umfasst die sichere Anwendungsentwicklung eine Reihe von Maßnahmen, darunter die Anwendung sicherer Programmierpraktiken, die Einbeziehung zuverlässiger Authentifizierungs- und Zugriffskontrollmechanismen, die Verwendung von Verschlüsselung sowie die Umsetzung wirksamer Verfahren zum Schwachstellenmanagement [1]. Darüber hinaus ist die sichere Softwareentwicklung kein einmaliger, sondern ein fortlaufender Prozess, der regelmäßige Sicherheitsaudits und Penetrationstests zur Ermittlung und Behebung potenzieller Schwachstellen umfasst.

Es ist jedoch unklar, ob dieser Ansatz für KI-basierte Anwendungen möglich und sinnvoll ist und ob er in der Praxis auch so umgesetzt wird. Ziel dieses Beitrags ist es daher, herauszufinden, welche Möglichkeiten es für die sichere Entwicklung von KI-Diensten gibt und ob diese in der Praxis eingehalten werden.

2. Übersicht von Merkmalen klassischer und KI-basierter Softwareentwicklung

Die klassische Softwareentwicklung und die Entwicklung von KI-Anwendungen haben viele Gemeinsamkeiten, weisen aber auch einige signifikante Unterschiede auf, insbesondere in Bezug auf Komplexität und Sicherheitsaspekte.

2.1. Gemeinsamkeiten

Sowohl in der klassischen als auch in der KI-basierten Softwareentwicklung sollte ein genereller Softwareentwicklungszyklus befolgt werden, der in der Regel Phasen wie Anforderungserfassung, Entwurf, Implementierung, Tests und Wartung umfasst [2–4]. Zudem ist in beiden Entwicklungsvarianten die Zusammenarbeit mit Entwicklerteams, Projektmanagern und Interessengruppen notwendig, um qualitativ hochwertige Softwareprodukte zu entwickeln, die den Anforderungen der Benutzer*innen entsprechen. Schlussendlich sind Softwaretests unerlässlich, um sicherzustellen, dass die entwickelte Software robust, zuverlässig und – im Falle einer KI-basierten Anwendung – frei von Bias ist.

2.2. Unterschiede

Bei der KI-Entwicklung wird mit Algorithmen für maschinelles Lernen (ML) und neuronalen Netzen gearbeitet, die viel komplexer sind als herkömmlicher Softwarecode [2]. Zudem ist die Vorgehensweise anders als in der herkömmlichen Softwareentwicklung und umfasst beispielsweise die Phasen Training, Modellierung, Anwendung und Ableitung [5]. All dies erfordert ein tiefes Verständnis der statistischen Modellierung, der Datenanalyse und der Konzepte des maschinellen Lernens. Zudem werden KI-Anwendungen oft auf großen Datensätzen trainiert, was zu Datenschutz- und Sicherheitsbedenken führen kann, wenn die Daten sensible Informationen enthalten.

Wenn Angreifende die Kontrolle über die Trainingsdaten oder das Modell erlangt, können diese das KI-System möglicherweise zu ihrem Vorteil manipulieren [6–10].

Weiters erfordern KI-Anwendungen oft große Mengen an Rechenressourcen, einschließlich Hochleistungscomputercluster und Spezialhardware wie GPUs. Dies kann die KI-Entwicklung teurer und ressourcenintensiver machen als die herkömmliche Softwareentwicklung. Schließlich werden KI-Systeme häufig in sensiblen Bereichen wie medizinischen Diagnosen, Finanzgeschäften oder Ausbildung [11] eingesetzt, was bedeutet, dass Fehler oder Sicherheitslücken schwerwiegende Folgen haben können.

3. Sichere Entwicklung von KI-Systemen in der Praxis

Nach diesem groben Überblick soll eine detaillierte Einsicht in die sichere Entwicklung von KI-Systemen in der Praxis gegeben werden. Zunächst werden einige Sicherheitsbedenken erwähnt, bevor Möglichkeiten zur sicheren und vertrauenswürdigen Entwicklung von KI-Applikationen aufgelistet werden. Zum Schluss wird dann die praktische Sichtweise der Entwickler*innen betrachtet.

3.1. Risiken und Sicherheitsbedenken

Sowohl herkömmliche Software als auch KI-Anwendungen können durch Sicherheitsbedrohungen wie Hackerangriffe, Malware und Datenschutzverletzungen gefährdet sein. KI-Systeme können jedoch aufgrund ihrer Komplexität und ihrer datenbasierten Art zusätzliche Sicherheitsbedenken aufwerfen.

Gemäß Khisamova et al. [12] gibt es zwei wesentliche Probleme im Zusammenhang mit dem Einsatz von KI, die Sicherheitsrisiken und ethische Bedenken aufwerfen können, nämlich erstens die Sammlung, Analyse und Verarbeitung von Daten und zweitens die KI-Entscheidungen auf Basis allgemeiner Daten. Ein besonderes Problem ist dabei, dass oft die Vertraulichkeit von Informationen nicht gewährleistet werden kann, insbesondere dann, wenn Schwachstellen in der Software vorhanden sind. Solche Schwachstellen können entweder zu zufälligen Fehlern führen oder böswilligen Akteuren die Möglichkeit geben, sensible Daten abzugreifen oder die Trainingsdaten bzw. das Modell zu manipulieren, um Bias einzuführen oder das System ungenauer zu machen [5, 7]. Dies kann in Bereichen wie der medizinischen Diagnose oder der Strafjustiz schwerwiegende Folgen haben. KI-Systeme können auch anfällig für Angriffe sein, bei denen ein Angreifer absichtlich Eingabedaten manipuliert, um das System zu falschen Entscheidungen zu veranlassen [7].

Li und Zhang [11] gehen auf etwaige Probleme genauer ein und identifizieren drei potenzielle Bereiche – Sicherheitsprobleme, Datenschutzprobleme und ethische Probleme. Unter Sicherheitsprobleme fallen dabei solche, die durch den Missbrauch der Technologie selbst entstehen, solche, die durch technische Fehler entstehen und solche, die durch die KI selbst entstehen, sollte die KI Bewusstsein erlangen. Datenschutzbedenken können im Bereich der Datenerfassung, Datennutzung in der Cloud oder bei der Wissensextraktion auftreten. Als ethische Probleme wird aufgezeigt, dass Roboter sozialen Regeln nicht folgen können, dass die Rolle von Robotern nicht klar definiert ist und ob es ethisch vertretbar ist, einen Roboter zu zerstören, der ein Bewusstsein entwickelt hat.

Dilmanghani et al. [5] geben schließlich eine genaue Übersicht darüber, welche Phase der KI-Entwicklung anfällig auf welchen Angriffsvektor ist. So ist die Trainingsphase besonders anfällig für Datenverluste, Bias in Daten und Data Poisoning, wohingegen das Modell selbst eine hohe Anfälligkeit auf Datenverluste und Modell-Extraktionsangriffe hat. Die Anwendungsphase ist angreifbar für Evasion-Angriffe, während die Ableitungsphase schließlich wieder besonders anfällig für Datenverluste ist.

Um die zuvor diskutierten Sicherheitsbedenken zu entkräften, müssen Entwickler von KI-Anwendungen ihre Systeme sorgfältig entwerfen und testen, um sicherzustellen, dass sie robust, genau und sicher sind. Zu diesem Zweck können Techniken wie Data Augmentation, Modellregularisierung und Adversarial Training eingesetzt werden, um die Genauigkeit und Widerstandsfähigkeit des KI-Systems zu verbessern [9]. Die Entwickler müssen auch die Auswirkungen der Daten, die sie zum Trainieren des Systems verwenden, auf den Datenschutz und die Sicherheit sorgfältig berücksichtigen und sicherstellen, dass das System vor potenziellen Angriffen geschützt ist.

Weitere Lösungen umfassen gemäß Li und Zhang [11] ein vermehrter Fokus auf Sicherheit, Datenschutz und Ethik in der Forschung, inklusive der Einbettung ethischer Regeln in die KI-Entwicklung, bessere Transparenz und Erklärbarkeit von KI-Systemen und die Verbesserung der Sicherheit und Robustheit der KI-Systeme. Zusätzlich sollte die KI-Entwicklung stärker reguliert und ein Fokus auf eine Verbesserung des Datenschutzes gelegt werden.

Das nächste Kapitel befasst sich daher mit Möglichkeiten zur sicheren und vertrauenswürdigen Entwicklung von KI-Applikationen.

3.2. Sichere und vertrauenswürdige Entwicklung von KI-Applikationen

Gemäß Wickramasinghe et al. [13] besteht der Lebenszyklus aus KI-Systemen aus den folgenden Phasen:

- *Initiierung*: Identifikation der Problemstellung und der Arten der Benutzer*innen sowie Erstellen des Konzeptvorschlages
- *Konzeptentwicklung und Planung*: Definition des Problemumfanges, Identifikation geeigneter Daten und anderer Ressourcen sowie Durchführbarkeitsstudie
- *Anforderungsanalyse*: Analyse der Benutzer*innenbedürfnisse und -anforderungen sowie Erstellen der Liste der funktionalen Anforderungen
- *Technisches Design:* Erstellen des Systementwurfs und Auflistung von Deliverables
- *Entwicklung und Tests:* Durchführen einer explorativen Datenanalyse, Erstellen der ML-Modelle und Abstimmung der Parameter, Modellvalidierung und -auswahl, sowie weitere Tests und Validierungen
- *Implementierung:* Einsatz des Systems in einer realen Umgebung und Schulung der Anwender*innen
- *Bereitstellung:* Durchführen von Leistungsüberwachungen und qualitativer Bewertung
- *Optimierung:* Verfeinerung des Modells, Interpretation des Modells und interaktive Visualisierungen

Wie zu sehen ist, sind in dem Lebenszyklus keine Sicherheitsanforderungen vorhanden, weswegen Galinkin [4] einen Responsible AI Development Lifecycle vorschlägt, um die Sicherheit von KI-Applikationen zu verbessern. Dieser Responsible AI Development Lifecycle ist an den klassischen Software Development Lifecycle angelehnt und besteht aus den Phasen Planung und Review, Design Review, Schadensmodellierung, Penetrationstests sowie Vorfallsmanagement.

- *Planung und Review* soll bei der Entwicklung neuer und Anpassung bestehender Systeme unterstützen, indem frühere Erkenntnisse in Betracht gezogen und die ersten Entwicklungsschritte definiert werden. Weitere Schritte in dieser Phase sind die Planung der Geschäftskontinuität und die Dokumentation der Pläne und Erkenntnisse.
- *Design Review* dient dazu, den Gesamtentwurf des Systems festzulegen, inklusive Zweck, algorithmischen Komponenten, Datenquellen, Protokollierung und Prüfung sowie Vor- und Nachverarbeitung.
- *Schadensmodellierung* betrachtet Gefahren und zeigt Maßnahmen auf, die vor der Inbetriebnahme ergriffen werden sollten. Auf Basis der erhobenen möglichen Schwachstellen, der betroffenen Akteure und der möglichen Auswirkungen erfolgt dann eine Risikobewertung der ML-Anwendung.
- *Penetrationstests* sollten nach Fertigstellung des AI-Systems durchgeführt werden, um mögliche Sicherheitslücken aufzudecken. Der Penetrationstest sollte statische Code-Scans, dynamische Schwachstellen-Scans und skriptgesteuerte Angriffe umfassen und idealerweise soweit wie möglich automatisiert bei jeder Aktualisierung durchgeführt werden
- *Vorfallsmanagement* tritt im Falle eines Problems in Kraft. Nach der Ermittlung des Umfangs und Ausmaßes des Schadens wird der Plan zur Aufrechterhaltung der Geschäftskontinuität angewendet, um das Problem zu beheben.

Der Responsible AI Development Lifecycle bietet zwar einen guten Überblick darüber, was für Maßnahmen durchgeführt werden sollen, um eine sichere und verantwortungsvolle Entwicklung von KI-Systemen zu gewährleisten, ist jedoch sehr allgemein gehalten und beantwortet nicht die Frage, wie die Schritte konkret durchgeführt werden könnten. Dementsprechend werden in weiterer Folge konkrete Maßnahmen und existierende Probleme, aufgezeigt von Avin et al. [14], vorgestellt:

- *Red Team Exercises* können der Phase der „Penetrationstests" zugeordnet werden und sind in der klassischen Softwareentwicklung weit verbreitet, um die Sicherheit der Systeme von unabhängigen, externen Fachleuten überprüfen zu lassen und etwaige Vertrauensprobleme zu beseitigen. Im Bereich KI-Entwicklung ist das Red-Teaming naturgemäß eher datengesteuert, weswegen Fachleute in diesem Bereich fehlen [14].

- *Audit Trails* können entweder gesetzlich vorgeschrieben oder auf freiwilliger Basis eingesetzt werden. Die Ermöglichung von Audits muss bereits in der Phase des „Design Reviews" berücksichtigt werden, da sie eine hohe Dokumentationsanforderung darstellen. Hier besteht jedoch das Problem, dass, obwohl Vorgaben zur Dokumentation und zum Auditing in anderen Domänen bereits weit verbreitet ist, ebenjene für die KI-Entwicklung noch nicht standardisiert existieren, was den Einsatz von KI-Applikationen in sensiblen Bereichen zunehmend erschwert [14]. Erste Rahmenwerke inkludieren Richtlinien zur Dokumentation bestimmter Merkmale von KI-Modellen und Vorschläge zur kontinuierlichen Protokollierung samt genauer Dokumentation der zum Trainieren der Modelle verwendeten Trainingsdaten und Ergebnisse.

- *Interpretierbarkeit und Erklärbarkeit* sind ebenfalls wesentliche Aspekte, um den "Blackbox"-Charakter von KI-Systemen aufzuweichen und die Gewährleistung der Sicherheit, Verantwortlichkeit und Fairness von KI-Systemen zu vereinfachen. Dieser Punkt sollte bereits in der Phase „Planung und Review" betrachtet werden, um zu definieren, ob der Einsatz von „Explainable AI" (XAI) Techniken nötig und sinnvoll ist. Gemäß Avin et al. [14] sollen solche Methoden einerseits dem Endbenutzer das Verständnis ermöglichen, wie ein Modell eine Ausgabe erzeugt, und andererseits in der Erklärung dem Modell treu bleiben und das zugrundeliegende Verhalten genau wiedergeben. Ergänzend zur Interpretierbarkeit und Erklärbarkeit ermöglicht die Reproduzierbarkeit externen Teams, ein KI-System nachzubauen und in Frage zu stellen, um die Behauptungen der Entwickler zu bestätigen. Initiativen wie das ACM Artifact Review and Badging und die ML Reproducibility Challenge bieten Anreize für die Reproduzierbarkeit im Forschungsumfeld [14].

- *Privacy-preserving Machine Learning*, auf Deutsch datenschutzfreundliches maschinelles Lernen, befasst sich damit, Datenschutzprobleme zu lösen, die beim maschinellen Lernen auftreten können und ist der Phase der „Schadensmodellierung" zuzuordnen. Solche Probleme umfassen beispielsweise den unbefugten Zugriff auf zum Trainieren der Modelle verwendeten Daten, die Verletzung der Privatsphäre oder der unbefugte Zugriff auf das trainierte Modell selbst [14]. Mögliche Techniken sind hierbei [14]:

    - *Föderierte Lerntechniken*, die das zentralisierte Training eines Modells mit dezentralisierten Daten ermöglichen, ohne dass die Rohdaten jemals das Quellgerät verlassen

    - *Differentielle Datenschutztechniken*, die den Entwicklungsprozess so anpassen, dass die trainierten Modelle aussagekräftige statistische Muster auf Bevölkerungsebene beibehalten, aber das Risiko des Rückschlusses auf Informationen über Einzelpersonen verringern, und

    - *Verschlüsselte Berechnungen*, die es Datenbesitzern und Modellentwicklern ermöglichen, Modelle zu trainieren, ohne dass eine der beiden Seiten Zugang zu den Informationen der anderen Seite hat.

  Es gibt bereits einige Open-Source-Implementierungsprojekte, wie Opacus von PyTorch, Tensorflow Privacy, FedAI, PySyft, Flower und OpenFL, aber kaum standardisierte Softwarebibliotheken, weswegen der Bekanntheitsgrad unter Entwickler*innen eher gering ist [14].

- *Bias und Sicherheitsmängel* treten vermehrt auf, da Schwachstellen und Risiken aufgrund der Komplexität von KI-Systemen vor ihrer Veröffentlichung oftmals nicht erkannt werden. In der klassischen Softwareentwicklung werden erkannte Schwachstellen schon lange veröffentlicht (z.B. CVE), in der KI-Entwicklung gibt es bis dato nichts Vergleichbares. Gemäß Avin et al. [14] stammt ein Großteil des Wissens über KI-Schäden von Forschern und Enthüllungsjournalisten, die nur begrenzten Zugang zu den von ihnen untersuchten KI-Systemen haben und oft antagonistische Beziehungen zu den Entwicklern unterhalten, deren Schäden sie aufdecken. Dieser Bereich ist der Phase der „Schadensmodellierung" zuzuordnen.

- *Meldung von KI-Vorfällen* ist wichtig, um zu verhindern, dass sich theoretische Risiken in tatsächliche Schäden verwandeln, und kann der Phase „Incident Response" zugeordnet werden. Aus Angst vor Reputationsschäden werden Vorfälle jedoch selten gemeldet, weswegen Anreize zur Weitergabe erforderlich sind. Beispielsweise könnte die Weitergabe gesetzlich vorgegeben werden oder Möglichkeiten zur anonymen Weitergabe an eine vertrauenswürdige Drittpartei (analog zu einem CERT) angedacht werden. Die Partnership on AI experimentiert mit einer solchen Plattform durch ihre AI Incident Database, die Informationen über KI-Vorfälle sowohl aus öffentlichen Quellen als auch aus Entwicklerberichten zusammenstellt, und das Center for Security and Emerging Technologies entwickelte eine Taxonomie mit drei Kategorien (Spezifikation, Robustheit und Sicherheit) auf der Grundlage gemeldeter Vorfälle, wobei mehr als 100 Vorfälle als Beispiele für jede Kategorie dienen. [14]

Dilmanghani et al. [5] gehen schließlich technisch weiter ins Detail und schlagen konkrete Gegen- und Abwehrmaßnahmen auf bekannte Angriffe vor, welche in jeder Phase des Responsible Development Lifecycles in Betracht gezogen werden sollten.

- Maßnahmen gegen Datenschutzverletzungen sind Anonymisierung, De-Identifizierung und Techniken zur Verbesserung der Privatsphäre (PET). Oftmals reicht es schon, sensible Informationen zu maskieren, um die Privatsphäre und Datensicherheit zu gewährleisten, beispielsweise durch die Anwendung von k-anonymity Maßnahmen. In sensibleren Bereichen sollte die Anwendung von PET-Techniken in Betracht gezogen werden. Ein Beispiel ist das OPen ALgorithms (OPAL) Projekt, welches Daten durch Zugangsprotokolle und Aggregationsverfahren schützt und es Forschern erlaubt, die Algorithmen, die mit den Daten trainiert werden sollen, zu übermitteln, anstatt die Daten aus der Hand zu geben. [5]

- Bias in Daten können mit verschiedenen statistischen Metriken erkannt werden, wie Mittelwertdifferenz, Restdifferenz oder normalisierter Information. Weitere Möglichkeiten sind die Verwendung von Lösungsansätzen wie optimierte Vorauswahl, das Lernen fairer Repräsentationen oder Adversarial Debiasing oder von fertigen Toolboxen wie Lime oder FairML. [5]

- Data Poisoning, also vergiftete Daten, können prinzipiell durch die Identifizierung von Ausreißern und Anomalien erkannt werden, außer wenn der Angreifer die Datenverteilung kennt und so die Anomalieerkennung umgeht [5].

- Modell-Extraktionsangriffe können durch die PATE-Technik, also durch das Trainieren mehrerer Modelle unter der Verwendung von verschiedenen Partitionen von Trainingsdaten für jedes Modell verhindert werden [5].

- Evasion-Angriffe führen zu Fehlklassifizierung durch kleine Änderungen, weswegen ein einfacher Abwehrmechanismus darin besteht, sicherzustellen, dass eine geringfügige Änderung der Eingabe das Ergebnis nicht signifikant verändern kann [5].

Zusammenfassend lässt sich sagen, dass bei der Entwicklung einer KI-Anwendung idealerweise der Lebenszyklus der verantwortungsvollen KI-Entwicklung befolgt werden sollte und darüber hinaus konkrete Maßnahmen zur Verteidigung gegen bekannte Angriffe auf die Modelle und Daten ergriffen werden sollten. Im nächsten Kapitel wird daher die Perspektive der Praktiker*innen vorgestellt.

3.3. Sichtweisen der Entwickler*innen

In 2020 wurde von Kumar et al. [15] eine Studie zur Industrieperspektive von 28 Organisationen im Bereich Adversariales maschinelles Lernen durchgeführt. Zu diesem Zwecke wurden 56 Fachleuten für maschinelles Lernen befragt, die für die Erstellung und Absicherung von ML-Modellen in ihrer jeweiligen Organisation zuständig sind. Die Ergebnisse dieser Studie waren ernüchternd und zeigen, dass die meisten der befragten Unternehmen weder über Tools noch über Know-How im Bereich der Absicherung der ML-Systeme verfügen und sich eher auf traditionelle Sicherheit und nicht auf KI-Sicherheit konzentrieren. Konkrete Probleme in bei der Entwicklung sind insbesondere eine geringe Kenntnis von möglichen Angriffen und spärliche Kenntnis von ML-spezifischen, sicheren Programmierverfahren. Weitere Probleme umfassen eine geringe Kenntnis der statischen und dynamischen Analyse von ML-Systemen, der Durchführung von Auditing und Logging sowie der Erkennung und Überwachung von ML-Systemen. In der Ausrollphase umfassen Probleme insbesondere das Fehlen von automatisierten Tests auf gegnerische Angriffe sowie ein geringes Wissen im Bereich Red Teaming. Im Falle eines konkreten Angriffes sehen die Autoren Verbesserungsbedarf bei dem Aufspüren und Bewerten von ML-Schwachstellen sowie bei der Reaktion auf Vorfälle, inklusive forensischer Maßnahmen und Wiederherstellung.

Darauf aufbauend wurde im Jahr 2021 von Boenisch et al. [16] eine Studie zum Sicherheits- und Datenschutzbewusstsein von insgesamt 83 Data Scientists veröffentlicht. Der Fragebogen umfasste 25 Fragen zu den Bereichen Demographie, Daten und Sensibilität, ML-Sicherheit, Angriffe, ML Datenschutz- und Sicherheitspraktiken sowie DSGVO-Kenntnis unterteilt. Die Ergebnisse zeigen auch hier, dass das Bewusstsein der befragten ML-Praktiker*innen für Bedrohungen sowie für ML-Sicherheits- und Datenschutzpraktiken relativ gering ist, was daran liegen kann. dass die meisten ML-Praktiker*innen keine akademische Ausbildung in ML Sicherheit und Datenschutz hatten. Konkret besteht wenig Vertrautheit mit Schutzstrategien und bekannten ML-Sicherheits- und Datenschutzbibliotheken.

Auch die Studie von Bieringer et al. [17] aus dem Jahr 2022 über die mentalen Modelle von Industriepraktiker*innen in Bezug auf das gegnerische maschinelle Lernen, in der 15 Praktiker*innen befragt wurden zeigte ähnliche Ergebnisse. So wurde festgestellt, dass die Unterscheidung zwischen ML-Bedrohungen und traditionellen Bedrohungen für die Praktiker*innen oftmals nicht klar ist und dass ML-spezifische Bedrohungen oft als weniger relevant als traditionelle Bedrohungen angesehen werden, obwohl durchaus praktische Erfahrung mit (erfolgreichen) Angriffen auf ML-Modelle besteht. Dementsprechend umfassen die Empfehlungen insbesondere die Aufklärung und Stärkung des Bewusstseins für gegnerische ML-Angriffe, die Einbindung von Maßnahmen zur Verhinderung von gegnerischem ML in Arbeitsabläufe und die Erstellung geeigneter regulatorischer und standardisierter Rahmenwerke in diesem Bereich.

4. Fazit

Zusammenfassend lässt sich sagen, dass die sichere Entwicklung von KI-Anwendungen von entscheidender Bedeutung ist, um die Anwendung und die Daten der Benutzer*innen vor Sicherheitsverletzungen zu schützen. Da sich KI weiterentwickelt und in unserem täglichen Leben immer häufiger vorkommt, ist es unerlässlich, der Sicherheit während des gesamten Entwicklungsprozesses Priorität einzuräumen. Durch die Implementierung von sicheren Programmierpraktiken, Zugriffskontrollen und Schwachstellenmanagementverfahren kann die Entwicklung von KI-Anwendungen sicherer und zuverlässiger gestaltet werden. Dies gestaltet sich jedoch in der Praxis schwierig, da etablierte Techniken, wie Red Teaming oder Auditing nicht eins zu eins in die KI-Entwicklung übernommen werden können, und neue Probleme, wie die der Erklärbarkeit und Interpretierbarkeit auftreten. Diese Sichtweise wird von KI-Entwicklern in der Praxis bestätigt, weswegen starker Forschungsbedarf im Bereich sicherer Entwicklung von KI-Systemen besteht.

Literatur

[1] R. A. Khan, S. U. Khan, H. U. Khan und M. Ilyas, „Systematic Literature Review on Security Risks and its Practices in Secure Software Development", *IEEE Access*, Jg. 10, S. 5456–5481, 2022, doi: 10.1109/ACCESS.2022.3140181.

[2] F. A. Batarseh, R. Mohod, A. Kumar und J. Bui, „The application of artificial intelligence in software engineering" in *Data Democracy*, Elsevier, 2020, S. 179–232, doi: 10.1016/B978-0-12-818366-3.00010-1.

[3] M. Howard und S. Lipner, *The security development lifecycle: SDL, a process for developing demonstrably more secure software*. Redmond Wash.: Microsoft Press, 2006.

[4] E. Galinkin, „Towards a Responsible AI Development Lifecycle: Lessons From Information Security", 6. März 2022. [Online]. Verfügbar unter: http://arxiv.org/pdf/2203.02958v1.

[5] S. Dilmaghani, M. R. Brust, G. Danoy, N. Cassagnes, J. Pecero und P. Bouvry, „Privacy and Security of Big Data in AI Systems: A Research and Standards Perspective" in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, S. 5737–5743, doi: 10.1109/BigData47090.2019.9006283.

[6] Duygu Sinanc Terzi, Ramazan Terzi und Seref Sagiroglu, „A survey on security and privacy issues in big data".

[7] M. Rigaki und S. Garcia, „A Survey of Privacy Attacks in Machine Learning", 15. Juli 2020. [Online]. Verfügbar unter: http://arxiv.org/pdf/2007.07646v2.

[8] Haina Ye, Xinzhou Cheng, Mingqiang Yuan, Lexi Xu, Jie Gao und Chen Cheng, „A survey of security and privacy in big data".

[9] Gayatri Sravanthi Kuntla, „Security and privacy in machine learning. A survey".

[10] R. Bao, Z. Chen und M. S. Obaidat, „Challenges and techniques in Big data security and privacy: A review", *Security and Privacy*, Jg. 1, Nr. 4, e13, 2018, doi: 10.1002/spy2.13.

[11] X. Li und T. Zhang, „An exploration on artificial intelligence application: From security, privacy and ethic perspective" in *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, China, 2017, S. 416–420, doi: 10.1109/ICCCBDA.2017.7951949.

[12] K. Zarina I., B. Ildar R. und S. Elina L., „Artificial Intelligence and Problems of Ensuring Cyber Security", 2020, doi: 10.5281/zenodo.3709267.

[13] C. S. Wickramasinghe, D. L. Marino, J. Grandio und M. Manic, „Trustworthy AI Development Guidelines for Human System Interaction" in *2020 13th International Conference on Human System Interaction (HSI): Tokyo, Japan, 06-08 June, 2020 : online proceedings*, S. Muramatsu, Hg., Piscataway, NJ: IEEE, 2020, S. 130–136, doi: 10.1109/HSI49210.2020.9142644.

[14] S. Avin *et al.*, „Filling gaps in trustworthy development of AI" (eng), *Science (New York, N.Y.)*, Jg. 374, Nr. 6573, S. 1327–1329, 2021, doi: 10.1126/science.abi7176.

[15] R. S. S. Kumar *et al.*, „Adversarial Machine Learning -- Industry Perspectives", 2020.

[16] F. Boenisch, V. Battis, N. Buchmann und M. Poikela, „"I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners" in *MuC '21: Mensch und Computer 2021*, Ingolstadt Germany, 2021, S. 520–546, doi: 10.1145/3473856.3473869.

[17] L. Bieringer, K. Grosse, M. Backes, B. Biggio und K. Krombholz, „Industrial practitioners' mental models of adversarial machine learning", *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, S. 97–116, 2022. [Online]. Verfügbar unter: https://www.usenix.org/conference/soups2022/presentation/bieringer