A-SIT

Zentrum für sichere Informationstechnologie - Austria

Sicherheits- und Privatsphäre-Aspekte von KI-basierten E-Government-Diensten

Sicherheits- und Privatsphäre-Aspekte von KI-basierten E-Government-Diensten

Autor:
Bianca Danczul
Mail: bianca.danczul@iaik.tugraz.at
Datum: 31.07.2022

Abstract/Zusammenfassung:
Der öffentliche Sektor setzt zunehmend auf KI-basierte E-Government-Dienste, da diese den Zeitaufwand und die Kosten auf Seiten der Behörden reduzieren, während die Zufriedenheit der Bürger*innen erhöht wird. Solche Dienste bergen aber naturgemäß auch Risiken, vor allem, wenn Persönlich identifizierbare Information (PII) verarbeitet werden. Dementsprechend ist es wichtig, dass sowohl bei deren Entwicklung als auch beim Einsatz Sicherheits- und Privatsphäre-Anforderungen betrachtet und eingehalten werden, da andernfalls die Glaubwürdigkeit und Vertrauenswürdigkeit der Behörden in Frage gestellt werden könnte.

In der klassischen Softwareentwicklung gibt es zur Absicherung der Dienste und zu deren Sicherheitsprüfung bereits eine Vielzahl von etablierten Softwareentwicklungs- und Sicherheitstest-Frameworks. Da es sich bei KI jedoch um ein relativ neues und umfangreiches Gebiet mit verschiedenen Unterthemen handelt, ist nicht klar, wie die Privatsphäre und Sicherheit KI-basierter Anwendungen in der Praxis gewährleistet werden kann.

Ziel der vorliegenden Studie ist es, zu analysieren, welche Möglichkeiten für die Verifikation der Sicherheits- und Privatsphäre von KI-basierten E-Government-Diensten existieren. Im Zuge der Studie sollen zudem etwaige Probleme und Forschungslücken aufgezeigt werden.

Inhaltsverzeichnis

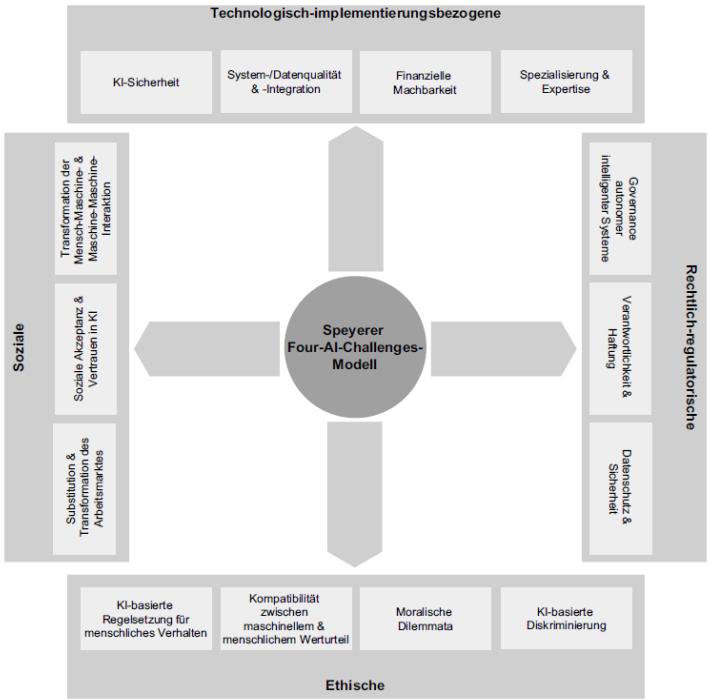
1. Einleitung

Die Idee von intelligenten Maschinen, so genannten Robotern, gab es schon lange, bevor der Begriff der künstlichen Intelligenz (KI) im Jahr 1956 offiziell eingeführt wurde. So dachte Aristoteles bereits im dritten Jahrhundert v. Chr. über die Möglichkeit nach, Sklaven durch Automatisierung zu ersetzen, Leonardo da Vinci entwarf um 1495 einen funktionstüchtigen, humanoiden, mechanischen Ritter, und die von Grey Walter entworfene "Schildkröte" aus dem Jahr 1948 kann als erster autonomer mobiler Roboter angesehen werden. Mit der Erfindung des Computers ergaben sich schließlich weitere Anwendungsgebiete, wie das Lösen von Rätseln und Spielen, der Einsatz von Expertensystemen oder automatisierten Planungssystemen bis hin zur automatisierten Entscheidungs- oder Wissensfindung. [1]

Um die Vorteile von KI, wie die Automatisierung von Arbeitsprozessen, die Verringerung des Verwaltungsaufwands oder die Verbesserung der Entscheidungsfindung und Servicequalität [2], auch in der öffentlichen Verwaltung nutzen zu können, wurden sowohl in Österreich [3–5] als auch in der EU [6–14] eine Vielzahl von Initiativen und Strategiedokumenten veröffentlicht, welche den Einsatz von KI forcieren und fördern sollen. Der Erfolg dieser Initiativen zeigt sich durch den immer stärkeren, EU-weiten Einsatz von KI-Systemen, vor allem in den Bereichen Chatbots und digitale Assistenten, Datenvisualisierung und Zukunftsforschung, Expertensysteme und algorithmische Entscheidungsfindung, Bildverarbeitung und Identitätserkennung, Verarbeitung natürlicher Sprache und Text Mining, Maschinelles Lernen und Deep Learning, Prozessautomatisierung und Kognitive Robotik, KI-gestütztes Wissensmanagement, Sicherheitsanalyse und Threat Intelligence sowie Audioverarbeitung [15].

In der österreichischen Verwaltung soll gemäß der Strategie KI nur im Einklang mit den europäischen Grundwerten, dem europäischen Rechtsrahmen und den Grund- und Menschenrechten eingesetzt werden, weshalb jedes KI-System folgende drei Grundprinzipien erfüllen muss, um als vertrauenswürdig zu gelten [3]:

- Es muss rechtmäßig sein und dabei alle bestehenden Gesetze und Regulierungen respektieren;
- Es muss ethische Prinzipien und Werte wie Gleichbehandlung und Fairness respektieren und
- Es muss robust sein, sowohl im technischen Sinn als auch aus gesellschaftlicher Perspektive.



Da es sich beim Einsatz von KI im öffentlichen Sektor um ein noch sehr junges Forschungsgebiet mit vielen unbeantworteten Fragen handelt [16], ist die Erfüllung dieser drei Grundprinzipien mit Herausforderungen in den Dimensionen Regulatorik, Ethik, Soziales und technologische Umsetzung verbunden, die sich nach Wirtz und Weyerer [2] in Speyers Four-AI-Challenges-Modell darstellen lassen (siehe Abbildung 1). Zu den technologisch-implementierungsbezogenen Hürden gehören insbesondere die Bereiche KI-Sicherheit, System-/Datenqualität und Integration, finanzielle Machbarkeit sowie Spezialisierung und Expertise [2].

Abbildung 1: KI-bezogene Herausforderungen nach dem Speyerer Four-AI-Challenges-Modell, Quelle: [2, S. 40]

Nachdem auch die österreichische KI-Strategie die Gewährleistung von KI-Sicherheit als einen wichtigen Teilbereich der vertrauenswürdigen KI nennt [3], befasst sich die vorliegende Studie mit den technologisch-implementierungsbezogenen Hürden und dabei konkret mit der Sicherheits- und Privatsphäre von KI-basierten Diensten. Der Fokus liegt dabei einerseits auf den Bedrohungen für KI-Systeme und andererseits auf der Sicherheit während des Lebenszyklus einer KI-Applikation.

2. Bedrohungslandschaft

Je nach Anwendungsfeld gibt es verschiedene Herausforderungen und Bedrohungen in Bezug auf AI-Applikationen. Um die wichtigsten Anwendungsfelder im E-Government-Bereich abzudecken, soll im Folgenden ein kurzer Überblick über jene im Bereich Machine Learning / Deep Learning, inklusive Big Data, und – als Spezialbereich – Chatbots gegeben werden,

2.1. Machine Learning

Das Ziel von maschinellem Lernen ist es, in Daten Zusammenhänge und Muster zu erkennen und diese Erkenntnisse in weiterer Folge auf bislang unbekannte Daten anzuwenden, um Vorhersagen treffen zu können. Deep Learning ist als Teilgebiet des maschinellen Lernens dafür zuständig, komplexere Probleme zu lösen, hat aber den Nachteil einer fehlenden Transparenz und komplizierten Gestaltung.

Der allgemeine Ansatz zur Erstellung eines Modells für maschinelles Lernen besteht aus drei Phasen, der Trainingsphase, der Validierungsphase und der Testphase. Dazu werden die Daten zu Beginn in Trainings-, Validierungs-, und Testdaten geteilt [17]:

- In der Trainingsphase erlernt der ML-Algorithmus die Merkmale der zu analysierenden Daten in Bezug auf eine bestimmte Aufgabe.
- In der Validierungsphase wird die Leistung des Modells anhand eines Testdatensatzes validiert, der unabhängig von den Trainingsdaten sein muss, um die Generalisierung des Modells zu messen.
- In der Testphase wird mittels eines Testdatensatzes getestet, ob das Modell korrekte Vorhersagen macht.

Der Einsatz von Big Data kann maschinelles Lernen verbessern, da dadurch eine große Menge an relevanten Daten zur Verfügung gestellt werden können, was wiederum eine Voraussetzung für funktionierende Modelle darstellt. Umso wichtiger ist es, dass beim Sammeln und Verwenden dieser Daten auf eine ausreichende Sicherheit und Privatsphäre geachtet wird, welche durch verschiedene Bedrohungen gefährdet sein kann. Solche Bedrohungen sind gemäß Dilmaghani et at. [18] beispielsweise Data Breaches, Bias in Data, Data Poisoning, Model Extraction und Evasion.

2.1.1. Data Breaches

Big Data zeichnet sich insbesondere durch einen großen Umfang und ein schnelles Wachstum an Daten aus. Die Menge an Daten kann je nach Unternehmen und Anwendungsfall durchaus in den Zetabyte-Bereich gehen, weswegen diese Daten oftmals in der Cloud gespeichert werden [19]. Werden diese Daten unzureichend abgesichert, kann dies zu einer unbefugten Offenlegung vertraulicher oder sensibler Daten, einem sogenannten Data Breach, führen. Erfolgreiche Data Breaches betreffen alle Phasen eines ML-Modells. [18]

Schutzmaßnahmen vor Data Breaches und zur Wahrung der Privatsphäre bei Big Data lassen sich prinzipiell in drei Klassen einteilen: Anonymisierung, De-Identifizierung und Privacy-Enhancing Technologies (PET) [18].

- Anonymisierung bedeutet, dass der Datensatz keine identifizierbaren Informationen enthält und es somit keine Möglichkeit gibt, die Informationen mit identifizierbaren Informationen zu verknüpfen.
- De-identifizierung bedeutet, dass der Datensatz zwar keine identifizierbaren Informationen enthält, es aber eine Möglichkeit gibt, die Informationen mit identifizierbaren Informationen zu verknüpfen.

Bekannte Mechanismen sind hierbei Pseudonymisierung oder K-Anonymisierung.

- Technologien zum Schutz der Privatsphäre (Privacy-Enhancing Technologies - PETs) sind Technologien, die grundlegende Datenschutzprinzipien verkörpern, indem sie die Nutzung personenbezogener Daten minimieren, die Datensicherheit maximieren und den Benutzer*innen somit ermöglichen, sich vor einer Zurückverfolgung ihrer Aktivitäten zu schützen.

2.1.2. Bias in Data

Auch wenn Bias in Daten nicht per se als Sicherheitsproblem angesehen werden, kann deren Auftritt einerseits zu falschen Ergebnissen und andererseits zu großen Vertrauensproblemen in das KI-System selbst führen, was die Datenverfügbarkeit und Integrität negativ beeinflussen kann. Solche Bias in Daten können beispielsweise Stichprobenverzerrungen (also eine unausgewogene Darstellung der Stichproben), Algorithmenverzerrungen (also systematische Fehler im System), oder präjudizielle Verzerrungen (also eine inkorrekte Bewertung der Daten) sowie Messverzerrungen (also eine schlechte Messung des Ergebnisses) sein und sind insbesondere für die Trainingsphase relevant [18].

Möglichkeiten zur Erkennung von Bias in Daten ergeben sich durch den Einsatz verschiedener Metriken, wie Mittelwertdifferenz, Differenz der Residualwerte, Chancengleichheit, disparate Einflussfaktoren und normalisierte wechselseitige Information sowie optimiertes Pre-Processing, Ablehnung der Optionsklassifizierung, Erlernen fairer Repräsentationen und adversariales Debiasing [18].

2.1.3. Data Poisoning

Werden verfälschte Trainingsdaten in das Modell eingespeist, spricht man von Data Poisoning. Dieser Angriff zielt darauf ab, das Modell zu verfälschen oder es dazu zu zwischen, falsche Ergebnisse zu produzieren, mit dem Ziel, das Modell entweder unbrauchbar zu machen oder ein Backdoor einzubauen, um das System nach Belieben auszunutzen. Data Poisoning Angriffe sind insbesondere für die Trainingsphase relevant [18].

Ein gängiger Ansatz zur Erkennung vom Data Poisoning ist die Identifizierung von Ausreißern (d. h. die Erkennung von Anomalien), da davon auszugehen ist, dass die injizierten Daten einer anderen Datenverteilung folgen. Weitere Möglichkeiten sind die Analyse der Varianz in den jeweiligen Klassen oder die separate Analyse von neu hinzugefügten Trainingsdaten. [18]

2.1.4. Model Extraction

Ist es möglich, aus der Modellextraktion auf die Datensätze zu schließen, die zum Trainieren des Modells verwendet wurden, kann je nach Sensibilität der trainierten Daten eine erhebliche Verletzung der Privatsphäre und Vertraulichkeit durch die Offenlegung dieser sensiblen Informationen erfolgen. Solch eine Modellextraktion ist durch ein erfolgreiches Reverse-Engineering verschiedenster ML-Modelle, die auf Techniken wie logistischer Regression, linearer Klassifikator, Support-Vektor-Maschine oder neuronalen Netzen basieren, möglich, indem beispielsweise die Ein- und Ausgabepaare beobachtet werden. Model Extraxtion Angriffe sind insbesondere für die Validierungsphase relevant [18]

Eine Erkennung von Angriffen auf die Modellextraktion ist, die Verteilung von aufeinanderfolgenden API-Anfragen zu analysieren und mit normalem, gutartigem Verhalten zu Vergleichen. Zudem können solche Angriffe durch das Trainieren mehrerer Modelle unter Verwendung unterschiedlicher Partitionen von Trainingsdaten für jedes Modell oder durch die Begrenzung der Informationen über die Wahrscheinlichkeitsbewertung für jedes Modell abgewehrt werden.

2.1.5. Evasion

Evasion ist ein weit verbreiteter Angriff, bei dem das Ziel des Angreifers darin besteht, die Erkennung so zu verändern, dass die Systeme zu einer falschen Klassifizierung verleitet werden. Dazu werden "Adversarial Samples", also böswillig manipulierte Eingaben verwendet, die für einen Menschen korrekt aussehen, aber den Klassifikator aus dem Konzept bringen, indem die Eingabeproben so verändert werden, dass sie in die falsche Kategorie eingestuft werden. Diese Veränderung ist im Vergleich zu ursprünglicher Eingabe nur minimal, führt aber dennoch zu einer Fehlklassifizierung. Evasion Angriffe sind insbesondere für die Testphase relevant. [18]

Ein möglicher Schutzmechanismus besteht daher darin, sicherzustellen, dass eine kleine Änderung der Eingabe das Ergebnis nicht wesentlich verändern kann. Dazu kann das Modell beispielsweise direkt mit Adversarial Samples trainiert werden, indem echte Labels dazugefügt werden, oder man versucht, die unerwünschten Proben zu erkennen und aus dem Datensatz zu entfernen. [18]

2.1.6. Weitere Angriffe

Zusätzlich zu den zuvor beschriebenen Angriffen auf die in Machine Learning Modellen verwendeten Daten gibt es noch weitere Angriffe, wie Model Extraction Attacks, Feature Estimation Attacks, Model Memorization Attacks, Identification Attacks, Inference Attacks oder Linkage Attacks [20] oder Poisoning Attacks, Impersonate Attacks und Inversaion Attacks [21], deren Definition den Ausmaß dieser Arbeit jedoch sprengen würde.

2.2. Chatbots

Chatbots sind eine spezielle Art von KI-Services, die von Millionen Menschen täglich – oftmals unbewusst – eingesetzt werden und aufgrund deren spezifischen Aufbaus separat behandelt werden sollen. Ein Chatbot besteht prinzipiell aus vier Komponenten, dem Client-Modul, dem Kommunikationsmodul, dem Antwortgenerierungsmodul und dem Datenbankmodul [22]:

- Das Client-Modul ist der Teil des Chatbots, der mit den Anwender*innen interagiert.
- Das Kommunikationsmodul ist die Infrastruktur, die Nachrichten vom Client-Modul an das Antwortgenerierungsmodul und vom Antwortgenerierungsmodul an das Datenbankmodul übermittelt.
- Das Antwortgenerierungsmodul ist zuständig für das tatsächliche Verstehen der Nachricht und einer Generierung einer geeigneten Antwort.
- Im Datenbankmodul sind alle für eine Konversation relevanten Daten gespeichert.

Angriffe auf Chatbots können auf alle Module erfolgen und damit die Privatsphäre der Personen, die mit dem Chatbot kommunizieren, negativ beeinträchtigen.

2.2.1. Angriffe auf das Client-Modul

In einem Client-Modul wird neben dem Empfang der Benutzerdaten auch eine etwaige Authentifizierung und Spracherkennung durchgeführt. Mögliche Angriffe auf dieses Modul sind unbeabsichtigte Aktivierungsangriffe, Gefälschte Antworten, Angriffe auf die Zugangskontrolle und Unerwünschte Sprachmuster [22]:

- Unbeabsichtigte Aktivierungsangriffe treten dann auf, wenn ein sprachgesteuerter Assistent durch einen Sprachbefehl aktiviert wird, ohne dass dies die Intention des/der Benutzer*in war. So ein Fall kann entweder unbeabsichtigt auftreten, indem der Assistent durch ein ähnlich klingendes Wort aktiviert wird und somit ungewollt Gespräche aufzeichnet, oder bewusst, indem ein*e Angreiferin den Assistenten remote durch Einspielen der Weckphrase aktiviert. Ein erfolgreicher Angriff bewirkt dabei eine Verletzung der Vertraulichkeit und Privatsphäre des Opfers.
- Gefälschte Antworten können dann auftreten, wenn ein Skill des Chatbots bzw. des persönlichen Assistenten kompromittiert wurde. Der bösartige Skill kann dem/der Benutzer*in dann beispielsweise vortäuschen, dass er bereits beendet wurde, sammelt aber im Hintergrund weiterhin Daten und verletzt somit die Privatsphäre der Benutzer*innen.
- Angriffe auf die Zugangskontrolle treten auf, wenn sich bösartige Apps als legitime Apps tarnen und somit unerwünschte Berechtigungen erhalten oder Zugangsdaten abfragen können.
- Unerwünschte Sprachmuster haben dasselbe Ziel wie unbeabsichtigte Aktivierungsangriffe, mit dem Unterschied, dass der Sprachaktivierungsbefehl verschleiert wird, also für das Opfer nicht erkennbar ist. Die Auswirkungen sind jedoch dieselben, nämlich eine Verletzung der Vertraulichkeit und Privatsphäre bei erfolgreicher Ausführung.

2.2.2. Angriffe auf das Kommunikationsmodul

Im Kommunikationsmodul werden einerseits die Nachrichten vom Client-Modul zum Antwortgenerierungsmodul transportiert und andererseits Datenabfragen vom Antwortgenerierungsmodul an das Datenbankmodul erfüllt. Dementsprechend betreffen die möglichen Angriffe hier vor allem den Netzwerkstack, durch Wiretapping, Man-in-the-Middle (MitM) Angriffe oder Distributed Denial of Service (DDoS) Angriffe [22]:

- Wiretapping ist selbst dann eine Gefahr, wenn der Datenverkehr im Kommunikationsmodul vollständig verschlüsselt ist, da der*die Angreifer*in unter Umständen selbst aus scheinbar harmlosen Metadaten Informationen extrahieren kann. Solche Metadaten können durch den Einsatz von Packet Sniffern wie Wireshark abgefragt werden und die Paketgröße oder die Anzahl der übertragenen Bytes enthalten, was Rückschlüsse auf den verwendeten Befehl ermöglicht.
- MitM Angriffe können Nachrichten zwischen Client A und Client B abfangen und durch eigene bösartige Nachrichten des Gegners ersetzen, mit dem Ziel, den/die Benutzer*in zu provozieren, Spam zu verbreiten oder sonstige Falschinformationen zu verbreiten.
- DDos-Angriffe zielen darauf ab, den Chatbot an der Interaktion mit Benutzern zu hindern, indem sie den Server mit Anfragen überfluten, was den Unternehmen, die sich auf den Chatbot verlassen, erheblichen Schaden zufügen kann.

2.2.3. Angriffe auf das Antwortgenerierungsmodul

Das Antwortgenerierungsmodul ist in erster Linie für die Interpretation der Nachricht und die Erstellung einer angemessenen Antwort verantwortlich. Man unterscheidet mögliche Angriffe in domänenfremde Angriffe, fehlerhafte Textbeispiele, Angriffe auf Sprachmodelle, Umprogrammierung durch einen Angreifer oder Feedback-Engineering-Angriffe [22]:

- Domänenfremde Angriffe kommen insbesondere bei sehr spezifischen Chatbots vor. Das Ziel eines solchen Angriffes ist es, durch Brute-Force-Angriffe Bereiche zu finden, in denen es dem Chatbot an fundiertem Wissen handelt. Diese Bereiche können anschließend so ausgenutzt werden, dass Benutzer*innen dazu verleitet werden könnten, den Angreifer*innen persönliche Informationen preiszugeben.
- Fehlerhafte Textbeispiele können verwendet werden, um den Chatbot dazu zu bringen, auf Eingabetexte mit falschen Informationen zu antworten oder beleidigende Sprache zu verwenden.
- Angriffe auf Sprachmodelle betreffen insbesondere NLP-Systeme. Eine Methode, um die derzeitige Abhängigkeit der Chatbots von Sprachmodellen auszunutzen, ist die Entwicklung bösartiger Modelle, die das NLP-System auf sehr spezifische Weise zum Versagen bringen können
- Umprogrammierung durch einen Angreifer bedeutet, dass diese*r das Antwortgenerierungsmodul so umprogrammiert, dass es eine andere Aufgabe erfüllt, ohne jedoch die Modellparameter zu ändern.
- Feedback-Engineering-Angriffe machen sich die Fähigkeit des Antwortgenerierungsmoduls zunutze, aus dem Nutzerfeedback zu lernen, indem Feedback erzeugt wird, welches den Chatbot negativ beeinflusst.

2.2.4. Angriffe auf das Datenbankmodul

Das Datenbankmodul ermöglicht dem Chatbot, für das Gespräch relevante Informationen nachzuschlagen, indem beispielsweise ein Wissensgraph abgefragt wird. Angriffe auf das Datenbankmodul, wie SQL-Injection oder Angriffe auf den Wissensgraphen, könnten die Privatsphäre der Benutzer*innen gefährden und das Verhalten des Chatbots grundlegend verändern.

- Eine erfolgreiche SQL-Injection ermöglicht die Durchführung unerlaubter Operationen auf der Datenbank, wie die Änderung von Informationen oder die Rückgabe sensibler Daten.
- Angriffe auf den Wissensgraphen führen dazu, dass Trainingsdaten so verändert werden können, dass bestimmte, wichtige Fakten hinzugefügt oder entfernt werden, um Korrelationen abzuändern.

3. Sichere Entwicklung von KI-Applikationen

Bei der Entwicklung von KI-Applikationen sollte darauf geachtet werden, dass die im vorherigen Kapitel genannten Angriffsszenarien betrachtet und nach Möglichkeit vermieden werden. Es ist zu beachten, dass alle Daten, Trainings-, Test-, und Validierungsdaten sowie die Betriebsdaten geschützt werden, um den Verlust sensibler Daten zu vermeiden.

Dementsprechend sollten die Grundsätze, die zum Schutz von Daten in anderen Systemen verwendet werden, auch auf KI- und ML-Projekte angewendet werden, einschließlich Anonymisierung, Tokenisierung und Verschlüsselung. Hierzu sollte das Prinzip Security by Design beachtet werden und sowohl Penetration-Tests als auch Red Teaming durchgeführt werden.

Eine Möglichkeit, diese Grundsätze zu beachten, ist die Verwendung eines sicheren Entwicklungsprozesses sowohl für herkömmliche Softwareprojekte als auch für KI- und ML-basierte Systeme, eine weitere ist die Einführung einer AI-Governance. Beide Varianten sollen in weiterer Folge kurz vorgestellt werden.

3.1. Responsible AI Development Lifecycle

Eine Möglichkeit, diese Grundsätze anzuwenden, ist die Umwandlung des sicheren Software Development Lifecycles auf KI-Projekte, wie von Galinkin [23] vorgeschlagen. Dieser Responsible AI Development Lifecycle besteht aus den Phasen Planung und Review, Design Review, Schadensmodellierung, Penetrationstests sowie Incident Response und soll im folgenden Kapitel kurz erklärt werden [23].

3.1.1. Planung und Review

Diese Phase stellt den ersten und zugleich letzten Schritt des AI Development Lifecycles dar und soll bei der Entwicklung neuer Systeme 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. [23]

3.1.2. Design Review

In dieser Phase wird der Gesamtentwurf des Systems festgelegt, inklusive dessen Zweck, den algorithmischen Komponenten und den zu verwendenden Datenquellen sowie der Protokollierung und Prüfung, der Vor- und Nachverarbeitung und der internen Verarbeitung. [23]

3.1.3. Schadenmodellierung

In dieser Phase werden mögliche Gefahren betrachtet und Abhilfemaßnahmen aufgezeigt, die vor der Inbetriebnahme ergriffen werden sollten. Dazu sollten alle beteiligten Personen miteinbezogen werden, um alle potenziellen Schäden für die Benutzer*innen und externe Parteien zu ermitteln und definieren. [23]

Ein weiterer Punkt, der in dieser Phase betrachtet werden sollte, ist die Risikobewertung der ML-Anwendung. Hierbei müssen die möglichen Schwachstellen, die Akteure und mögliche Auswirkungen in Betracht gezogen werden. [24]

3.1.4. Penetrationstests

Nach Fertigstellung des AI-Systems sollten ein Penetrationstests 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. [25]

Das Hauptaugenmerk sollte einerseits auf möglichen Angriffen auf das Modell liegen, aber auch potenzielle Repräsentationsschäden in Betracht ziehen [23]. Außerdem sollten Benutzer*innen und deren Sicherheitsteams nach dem Einsatz der ML-Applikation regelmäßig Sicherheitsteams und Red Teamings durchführen [24].

3.1.5. Incident Response

Wenn das System einsatzbereit ist, tritt im Falle eines Problems diese Phase in Kraft. Hier wird nach der Ermittlung des Umfangs und Ausmaßes des Schadens der Plan zur Aufrechterhaltung der Geschäftskontinuität angewendet, um das Problem zu beheben. Ist das Problem erfolgreich behoben, sollten die Ergebnisse geprüft und die Abhilfemaßnahmen geplant werden. [23]

Zu beachten ist, dass hierbei der Fokus nicht nur auf Patches als Abhilfestrategie gelegt wird, sondern auch die Widerstandsfähigkeit der ML-Applikation erhöht wird [24].

3.2. AI Governance in der Entwicklung

Laato et al. [26] schlagen vor, Governance-Aspekte bereits bereits in der Phase der Implementierung des KI-Systems einzubeziehen. Sie behandeln dabei die Phasen des System Design, System Development und System Operation, ohne jedoch dediziert auf Informationssicherheitsaspekte einzugehen. Dennoch soll ein kurzer Überblick über deren Ansatz gegeben werden.

3.2.1. System Design

In der System Design Phase werden der Business Case, die Datenressourcen und die externe Umgebung betrachtet, um die Voraussetzungen abzuklären [26]:

- Der Business Case befasst sich einerseits mit der Betrachtung, ob Machine Learning für das Geschäftsfeld geeignet ist und andererseits mit den projektbezogenen Bedürfnissen.
- Die Betrachtung der Datenressourcen ist wichtig, um festzustellen, ob ausreichend Trainings-, Test-, und Validierungsdaten vorhanden sind, ob die Verwendung der Daten ethisch vertretbar ist und ob es überhaupt rechtlich erlaubt und technisch möglich ist, die Daten zu verarbeiten.
- Die Betrachtung der externen Umgebung befasst sich damit, die Grenzen des Systems festzustellen und zu definieren, was passiert, wenn die Vorhersagen des ML-Modells falsch sind. Zudem muss der Einfluss der zur Entwicklung verwendeten technischen Hilfsmittel betrachtet und die Einhaltung gesetzlicher Vorgaben sowie anderer Vorschriften und Richtlinien gewährleistet werden. Dies sollte in jedem Fall eine ausführliche Risikobewertung der ML-Applikation beinhalten [24].

Ist geplant, dass das System von den empfangenen Daten im Betrieb lernt und sich so weiterhin anpasst, muss gewährleistet sein, dass die Datenströme nicht von bösartigen Akteuren manipuliert oder verfälscht werden. Zu diesem Zwecke könnte der Einsatz einer DLP-Lösung angedacht werden, die mögliche Bedrohungen verhaltensbasiert analysiert. [25]

3.2.2. System Development

Die System Development Phase befasst sich einerseits mit der Erstellung der Datensets und andererseits mit der Erstellung der Modelle sowie der Durchführung von Systemtests [26]:

- Bei der Erstellung der Datensets muss auf eine ausreichende Versionierung geachtet werden, sodass nachvollzogen werden kann, welches Modell welche Datensätze im Training verwendet. Zudem ist es wichtig, dass die Daten vor Verwendung validiert werden, sodass eine ausreichende Qualität gewährleistet werden kann.
- Bei der Erstellung der Modelle ist ebenfalls eine nachvollziehbare Versionierung der verschiedenen Modelle notwendig. Außerdem sollte sichergestellt werden, dass die verwendeten Parameter und die Spezifika der einzelnen Algorithmen dokumentiert werden, um die Nachvollziehbarkeit zu gewährleisten.
- Systemtests sind in dieser Phase wichtig, um einen Überblick über die Modelle und deren Probleme zu erlangen und die die Datensätze zu testen.

3.2.3. System Operation

In der System Operation Phase wird die Systemprüfung, die Erklärbarkeit der Modelle und die automatisierte Überwachung behandelt [26]:

- In dieser Phase befasst sich die Systemprüfung damit, zu kontrollieren, wie gut das System in die Betriebsumgebung passt.
- Die Erklärbarkeit der Modelle kann insbesondere bei Black-Box-Systemen aufgrund deren Komplexität kaum gewährleistet werden. Die Interpretierbarkeit der Systeme ist dabei umso schwerer, je komplexer das ML-Modell ist. Dementsprechend können zur Verbesserung der Erklärbarkeit verschiedene Frameworks, wie SHAP oder LIME eingesetzt werden.
- Automatisierte Überwachung ist notwendig, um die Genauigkeit und Funktionalität der ML-Systeme zu gewährleisten. Hierbei werden Techniken zur Anomalieerkennung, wie ungewöhnliche Vorkommnisse und Erkennung von Bias, wie unfaire Vorhersagen des Modells, verwendet.

Benutzer*innen und deren Sicherheitsteams sollten beim Einsatz einer ML-Applikation ebenfalls regelmäßig Sicherheitsteams und Red Teamings durchführen [24].

4. Fazit

In der vorliegenden Studie wurden zwei Möglichkeiten diskutiert, die Entwicklung von KI-Applikationen sicherer zu gestalten. Während es zur Verwendung von KI zur Verifikation der Sicherheit von klassischen Softwareprojekten viel Forschung gibt, ist die Gewährleistung der Sicherheit von KI-Systemen noch am Anfang. Doch trotzdem oder gerade deswegen ist es wichtig, auch die Entwicklung von KI-Systemen sicherer zu gestalten, um Bedrohungen wie Data Poisoning, Bias in Daten, Data Breaches zu vermeiden oder zumindest zeitnah feststellen zu können.

Die vorgestellten Möglichkeiten stellen einen ersten Startpunkt dar, die Sicherheit von KI-Applikationen sowohl im privaten als auch im öffentlichen Sektor zu verbessern. Nichts desto trotz ist nach wie vor zukünftige Forschung insbesondere im Bereich Nachvollziehbarkeit und Auditierbarkeit der Systeme, Schutz der Systeme gegen Manipulation, Testen von KI-Systemen sowie Sicherheit von KI-Systemen im gesamten Lebenszyklus erforderlich, um sicherzustellen, dass die gesetzlichen und ethischen Vorgaben eingehalten werden können.

Literatur

[1] S. J. Russell und P. Norvig, *Artificial Intelligence: A Modern Approach,* 4. Aufl. Hoboken: Pearson, 2021.

[2] B. W. Wirtz und J. C. Weyerer, „Künstliche Intelligenz im öffentlichen Sektor: Anwendungen und Herausforderungen", *VM*, Jg. 25, Nr. 1, S. 37–44, 2019, doi: 10.5771/0947-9856-2019-1-37.

[3] BMK, *Strategie der Bundesregierung für Künstliche Intelligenz. Artificial Intelligence Mission Austria. 2030.: (AIM AT 2030).* [Online]. Verfügbar unter: https://www.bmdw.gv.at/Themen/Digitalisierung/Strategien/Kuenstliche-Intelligenz.html#:~:text=Um%20die%20Chancen%20von%20KI,(AIM%20AT%202030)%20entwickelt.

[4] BMK, *Strategie der Bundesregierung für Künstliche Intelligenz. Annex.* [Online]. Verfügbar unter: https://www.bmdw.gv.at/Themen/Digitalisierung/Strategien/Kuenstliche-Intelligenz.html#:~:text=Um%20die%20Chancen%20von%20KI,(AIM%20AT%202030)%20entwickelt.

[5] Stadt Wien, *Künstliche Intelligenz Strategie: Digitale Agenda Wien.* [Online]. Verfügbar unter: https://digitales.wien.gv.at/wp-content/uploads/sites/47/2019/09/StadtWien_KI-Strategiepapier.pdf.

[6] *MITTEILUNG DER KOMMISSION AN DAS EUROPÄISCHE PARLAMENT, DEN EUROPÄISCHEN RAT, DEN RAT, DEN EUROPÄISCHEN WIRTSCHAFTS- UND SOZIALAUSSCHUSS UND DEN AUSSCHUSS DER REGIONEN. Künstliche Intelligenz für Europa: COM(2018) 237 final*, 2018.

[7] *Anhang der Mitteilung Der Kommission An Das Europäische Parlament, Den Europäischen Rat, Den Rat, Den Europäischen Wirtschafts- und Sozial¬ausschuss und den Ausschuss der Regionen. Koordinierter Plan für künstliche Intelligenz.: COM(2018) 795 final Annex*, 2018.

[8] *WEISSBUCH. Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen: COM(2020) 65 final*, 2020.

[9] *Mitteilung Der Kommission An Das Europäische Parlament, Den Europäischen Rat, Den Rat, Den Europäischen Wirtschafts- und Sozial¬ausschuss und den Ausschuss der Regionen. Koordinierter Plan für künstliche Intelligenz.: COM(2018) 795 final*, 2018.

[10] *ANHÄNGE der MITTEILUNG DER KOMMISSION AN DAS EUROPÄISCHE PARLAMENT, DEN RAT, DEN EUROPÄISCHEN WIRTSCHAFTS- UND SOZIALAUSSCHUSS UND DEN AUSSCHUSS DER REGIONEN. Förderung eines europäischen Konzepts für künstliche Intelligenz: COM(2021) 205 final Annex*, 2021.

[11] *ANHÄNGE des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates ZUR FESTLEGUNG HARMONISIERTER VORSCHRIFTEN FÜR KÜNSTLICHE INTELLIGENZ (GESETZ ÜBER KÜNSTLICHE INTELLIGENZ) UND ZUR ÄNDERUNG BESTIMMTER RECHTSAKTE DER UNION: COM(2021) 206 final*, 2021.

[12] *MITTEILUNG DER KOMMISSION AN DAS EUROPÄISCHE PARLAMENT, DEN RAT, DEN EUROPÄISCHEN WIRTSCHAFTS- UND SOZIALAUSSCHUSS UND DEN AUSSCHUSS DER REGIONEN. Förderung eines europäischen Konzepts für künstliche Intelligenz: COM(2021) 205 final*, 2021.

[13] *Vorschlag für eine VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES ZUR FESTLEGUNG HARMONISIERTER VORSCHRIFTEN FÜR KÜNSTLICHE INTELLIGENZ (GESETZ ÜBER KÜNSTLICHE INTELLIGENZ) UND ZUR ÄNDERUNG BESTIMMTER RECHTSAKTE DER UNION: COM(2021) 206 final*, 2021.

[14] *Opportunities of Artificial Intelligence*, 2020.

[15] G. Misuraca und C. van Noordt, *AI Watch - artificial intelligence in public services: Overview of the use and impact of AI in public services in the EU*. Luxembourg: Publications Office of the European Union, 2020.

[16] B. W. Wirtz, J. C. Weyerer und C. Geyer, „Artificial Intelligence and the Public Sector—Applications and Challenges", *International Journal of Public Administration*, Jg. 42, Nr. 7, S. 596–615, 2019, doi: 10.1080/01900692.2018.1498103.

[17] E. D. Cristofaro, „A Critical Overview of Privacy in Machine Learning", *IEEE Secur. Privacy*, Jg. 19, Nr. 4, S. 19–27, 2021, doi: 10.1109/MSEC.2021.3076443.

[18] 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" (eng), *2019 IEEE International Conference on Big Data*, 2019, doi: 10.1109/BigData47090.2019.

[19] 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.

[20] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi und Z. Lin, „When Machine Learning Meets Privacy", *ACM Comput. Surv.*, Jg. 54, Nr. 2, S. 1–36, 2022, doi: 10.1145/3436755.

[21] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu und V. C. M. Leung, „A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View", *IEEE Access*, Jg. 6, S. 12103–12117, 2018, doi: 10.1109/ACCESS.2018.2805680.

[22] W. Ye und Q. Li, „Chatbot Security and Privacy in the Age of Personal Assistants" in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*, San Jose, CA, USA, 2020, S. 388–393, doi: 10.1109/SEC50012.2020.00057.

[23] Galinkin und Erick, „Towards a Responsible AI Development Lifecycle: Lessons From Information Security.", *AIES '22: 5th AAAI/ACM Conference*, 2022. [Online]. Verfügbar unter: https://arxiv.org/pdf/2203.02958.pdf

[24] A. Lohn und W. Hoffman, „Securing AI: How Traditional Vulnerability Disclosure Must Adapt", 2022.

[25] J. Burke, „Securing AI during the development process", *TechTarget*, 25. Jan. 2022, 2022. [Online]. Verfügbar unter: https://www.techtarget.com/searchenterpriseai/tip/Securing-AI-during-the-development-process. Zugriff am: 25. Juli 2022.

[26] S. Laato, T. Birkstedt, M. Mantymaki, M. Minkkinen und T. Mikkonen, „AI Governance in the System Development Life Cycle: Insights on Responsible Machine Learning Engineering", *1st Conference on AI Engineering - Software Engineering for AI (CAIN'22)*, S. 113–123, 2022, doi: 10.1145/3522664.3528598.