Beschaffung sicherer AI-Systeme – Herausforderungen & Praktischer Leitfaden

Peter Kieseberg, Laura Kaltenbrunner, Marlies Temper, Simon Tjoa, Institut für IT-Sicherheitsforschung, FH St. Pölten

Abstract. AI ist als Werkzeug nicht mehr wegzudenken und wird auch in Zukunft immer wichtiger werden. In diesem Vortrag beschäftigen wir uns mit wesentlichen Aspekten, die AI-Systeme in Bezug auf Sicherheitsanalysen anders machen als traditionelle Systeme und stellen unseren Beschaffungsleitfaden für sichere AI vor.

Keywords: AI, Beschaffung, Security

Beitrag im Rahmen des 16. Forschungsforums der österreichischen Fachhochschulen von 19.-20. April 2023 an der FH St. Pölten.

1 EINLEITUNG

Datengetrieben Anwendungen werden immer wichtiger und durchsetzen immer weitere Teile unseres täglichen Lebens. In den nächsten Jahren werden AI-basierte Systeme immer allgegenwärtiger werden und auch in alltäglichen Anwendungen eingesetzt werden [1]. Dabei zeitigen diese Systeme jedoch, neben all den durch sie entstehenden Chancen, einige Probleme sicherheitstechnischer Natur, speziell in Bezug auf fehlende Transparenz und das sog. Explainability-Problem. Speziell in kritischen Infrastrukturen kann dieses Emergenz-Problem große Sicherheitslücken entstehen lassen. Der AI-Act („Vorschlag für eine Verordnung zur Festlegung von harmonisierten Regeln über Künstliche Intelligenz“) [2] verkörpert daher die künftige Strategie der Europäischen Union in Bezug auf die Absicherung intelligenter Systeme, sowohl in Hinblick auf die Anwendung von AI in einem breiten Feld an Applikationen und mit speziellem Fokus auf Anwendungen im sicherheitskritischen Bereich. Allerdings ist bei viele populären Methoden der AI nicht klar, wie die im AI-Act gefordert Transparenz- und Sicherheitsanforderungen überhaupt gewährleistet werden können. Das Testen dieser Methoden auf Sicherheitslücken ist wesentlich komplexer als „klassischer“ Algorithmen, selbst wenn alle wichtigen Teile wie Trainingsdaten, Auswertungsdaten, Algorithmen und Modelle zur Verfügung stehen [3]. Dies wir jedoch künftig oftmals nicht der Fall sein: Vortrainierte Modelle, Modelltraining als Service, aber auch komplette API-angesteuerte Black-Box-Lösungen werden für viele EntwicklerInnen das Mittel der Wahl sein um niederschwellig AI nutzen zu können. Dabei wird nicht nur IT-Security zum Problem, sondern auch in den Daten (oftmals irrtümlich) festgeschriebene Vorurteile, der sog. „bias“. Daher werden diese Themen auch in der Beschaffung immer wesentlicher, speziell, da derzeit viele Anbieter AI-Systeme im kritischen Bereich anbieten, ohne dass sie Sicherheitsüberlegungen in ihre Produkte einfließen ließen.

Im vorliegenden Vortrag werden wir auf die wichtigsten Themen um das Thema Sicherheitstests in AI-Systemen eingehen und speziell auf die sichere Nutzung von AI durch Nicht-AI-ExpertInnen eingehen. Wir werden auch unseren iterativ verbesserten „Beschaffungsleitfaden für sichere AI-Systeme“ vorstellen, der auf der FUZZ-IEEE-Konferenz vorgestellt wurde [4].

2 METHODEN

Im Rahmen des FORTE-Projekts exploreAI hatten wir das Werkzeug der explorativen Szenarienanalyse genutzt um wichtige treibende Faktoren in der Entwicklung, speziell aber der Nutzung, von AI-basierten Systemen zu analysieren. Daraus entstanden mehrere Entwicklungsszenarien, auf deren Basis wir mögliche sicherheitsrelevante Entwicklungen extrahiert hatten. Durch eine Gap-Analyse konnten wichtige mögliche Forschungslücken identifiziert werden, auf deren Basis unter Nutzung unseres Know-Hows in den Bereichen IT-Sicherheit und AI wir unsere Ergebnisse erhielten.

2.1 Explorative Szenarienanalyse

Im Rahmen von Interviews aufbauend auf einer extensiven Literaturrecherche wurden ExpertInnen gefragt, welche Faktoren sie als die wichtigsten Treiber in der Entwicklung

und Nutzung von AI und AI-basierten Anwendungen sehen. Diese Faktoren wurden zu Szenarien kombiniert, wobei verschiedene als wahrscheinlich erachtete und zueinander kompatible Entwicklungspfade zu Zukunftsszenarien kombiniert wurden.

Wesentlich war auch die Betrachtung klassischer Vorgehensweisen im Bereich der Sicherheitstests, speziell Penetration-Tests und der dabei verwendeten Strategien und Techniken wie bspw. Fuzzy Testing. Dabei wurden klassische Techniken versuchsweise im Experiment auf AI-basierte Systeme angewandt und in Hinblick auf neue Herausforderungen untersucht.

Auf Basis dieser Szenarien wurden sinnvolle Anwendungen extrapoliert und in Hinblick auf neu entstehende Sicherheitsproblematiken untersucht. Diese wurden erfasst und analysiert, inwieweit die sicherheitsrelevanten Aspekte schon im Zuge der Beschaffung analysier werden müssen.

2.2 Leitfadengenerierung

Basierend auf den Gaps und möglichen Mitigationsstrategien zum Zeitpunkt der Beschaffung, wurde ein Leitfaden generiert. Ziel dieses Leitfadens ist dabei rein die Triage von angebotenen Produkten: Weder kann der Leitfaden entscheiden, ob ein System sicher ist, noch, ob der Einsatz von AI überhaupt die richtige Strategie für ein Problem ist. Er liefert aber ein Set an Fragen, die es ermöglich festzustellen ob (i) einem Produkt grundsätzliche wesentliche Sicherheitsüberlegungen zu Grunde liegen, (2) wie mit wesentlichen Themen wie Kontrolle über Daten und Modelle, Patching, und dergleichen umgegangen wird und (3) ob einem überhaupt eine geeignete Ansprechperson gegenübersitzt, die solcherart Fragen sinnvoll und korrekt beantworten kann. Damit kann der Beschaffungsprozess in der initialen Auswahlphase, aber auch einer späteren Evaluierungsphase deutlich straffer und performanter gestaltet und ungeeignete Systeme und/oder Ansprechpartner frühzeitig aus der Auswahl genommen werden.

3 Probleme klassischer Sicherheitstests in AI-basierten Systemen

In diesem Kapitel wollen wir kurz auf wesentliche Herausforderungen bei der Anwendung klassischer Sicherheitstests auf AI-basierte Systeme eingehen, wobei natürlich vom Umfang her nicht auf Details eingegangen werden kann.

Ein wesentliches Problem betrifft den Erkenntnisgewinn durch Testings selbst. Im Fall von Algorithmen, die dem Explainability-Problem unterliegen kann es extrem schwierig sein, selbst bei einem erfolgreichen Angriff nachzuvollziehen, wo genau die Schwachstelle liegt.

Ein wesentlicher Aspekt dabei ist auch das Problem der Datenkontrolle – da bei vielen Methoden der AI die Daten wesentliche Aspekte der Software kontrollieren und definieren, können diese aus den Sicherheitsbetrachtungen nicht weiter außen vor gelassen werden, sondern müssen ebenfalls mit betrachtet werden. Dabei ist zu klären, woher diese Daten (speziell Trainingsdaten) überhaupt stammen und wie vertrauenswürdig die Quellen sind, aber auch, wer überhaupt die Kontrolle über Daten und Modelle besitzt: Werden diese weiter gewartet, sind diese fix, wer führt die Updates durch, wer kontrolliert die Integration neuer Daten … .

Speziell problematisch sind dabei Methoden des Reinforcement Learnings, da diese

konstant weiterlernen – dies ermöglicht nicht nur sehr raffinierte Methoden des Data Poisonings, es führt auch zu sehr grundsätzlichen Problemen in der Durchführung von Security-Tests: Da sich das System beständig ändert, wie lange „gilt“ ein durchgeführter Test, d.h. wie lange sind die Aussagen eines durchgeführten Sicherheitstests überhaupt gültig und ab wann hat sich das System soweit verändert, dass neu getestet werden muss? Dies ist vor allem in Hinblick auf Regularien die beständiges testen und zertifizieren im Fall von (größeren) Änderungen verlangen von Bedeutung. Zusätzlich muss auch die Frage gestellt werden, ob und wie gewisse Tests wie Fuzzy-Testing direkten Einfluss auf das System selbst nehmen und u.U. als Teststrategie sinnlos, bzw. Schäden an Echtsystemen zeitigen können.

Neben diesen Problemen ist auch das Thema der Erkennung von Vorurteilen ein wichtiger Aspekt, der durch derzeitige Penetrationtest-Methoden kaum abgedeckt wird. Last, but not least, betrifft dies auch Aspekte der Data Preparation, speziell Effekte, die durch die Integration von Anonymisierungsalgorithmen und Data Cleansing, ungewollt in das System eingebracht werden und derzeit kaum berücksichtigt werden.

4 Der Beschaffungsleitfaden

Es ist davon auszugehen, dass die meisten Anwender*innen AI zukaufen werden, bewusst oder auch unbewusst in Form von fertigen Modellen, APIs oder einfach Systemkomponenten. Damit muss sich auch die Beschaffung selbst mit den neuen durch AI induzierten Sicherheitsproblematiken beschäftigen. Im Rahmen dieses Vortrags werden wir die neue Version unseres „Beschaffungsleitfadens sichere AI-Systeme“ vorstellen. Die Originalversion wurde zusammen mit österreichischen Behörden im Rahmen des FORTE-Projekts exploreAIA entwickelt, die im Rahmen dieses Vortrags erstmalig präsentierte Version stellt eine grundlegende Überarbeitung dar. Das Ziel dieses Leitfadens besteht darin, den Anwendenden eine Reihe wichtiger Fragestellungen zur Auswahl geeigneter AI-Systeme in die Hand zu geben und damit der Flut potenzieller Anbieter Herr zu werden. Damit ist er auch ein wichtiges Werkzeug, um schon vor der Beschaffung wesentliche sicherheitsrelevante Aspekte, wie bspw. Kontrolle über Daten und Modelle, oder Data Cleansing, zu diskutieren und in einer etwaigen Ausschreibung abdecken zu können. Zusätzlich ermöglicht er es auch, abzuschätzen, ob ein Gesprächspartner im Rahmen eines Beschaffungsvorgangs das notwendige Wissen mitbringt, bzw. ob das anbietende Unternehmen überhaupt eine Security-Strategie besitzt. Damit kann schon frühzeitig die Spreu vom Weizen getrennt und der Beschaffungsprozess wesentlich beschleunigt und abgesichert werden.

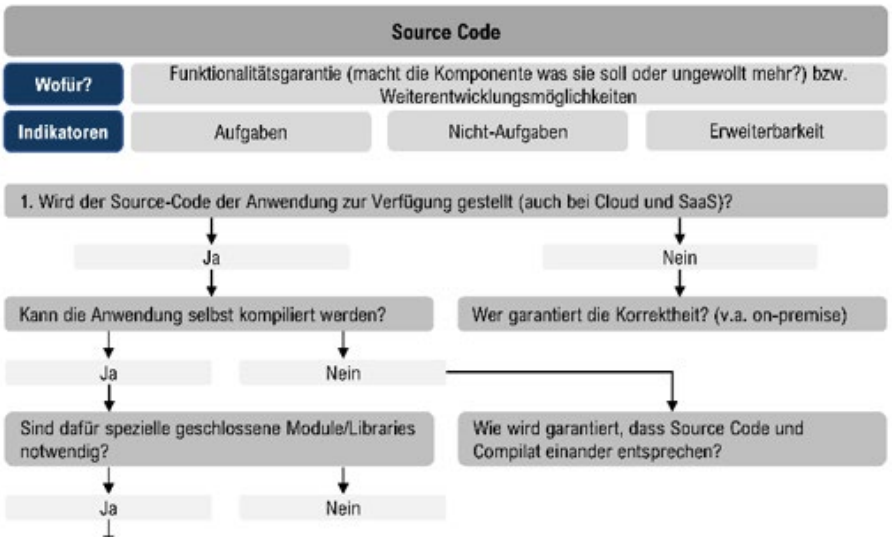

Abbildung 1. Beispiel aus dem Beschaffungsleitfaden

5 CONCLUSIO

Im vorliegenden Vortrag haben wir einen kurzen Abriss über die Problematiken von klassischen Sicherheitstests in datenbasierten AI-Systemen gegeben. Dieser zusätzliche Grad an Komplexität wird derzeit noch kaum berücksichtigt, wiewohl entsprechende Anforderungen gerade aus neuen Regularien wie dem AI-Act entstehen, die speziell im Bereich der High-Risk-AI sehr hohe Anforderungen an das Risikomanagement legen. Zusätzlich haben wir eine überarbeitete Version unseres Beschaffungsleitfadens für sichere AI vorgestellt, der das Thema der sicheren AI bereits in den Akt der Beschaffung einfließen lässt.

Der Leitfaden steht natürlich unentgeltlich und ohne finanzielle Interessen unter www.secureai.info zum Download zur Verfügung.

6 REFERENZEN

[1] Wang, X., Ren, X., Qiu, C., Cao, Y., Taleb, T. and Leung, V.C., 2020. Net-in-AI: a computing-power networking framework with adaptability, flexibility, and profitability for ubiquitous ai. IEEE Network, 35(1), pp.280-288.

[2] Helberger, N. and Diakopoulos, N., 2022. The European AI act and how it matters for research into AI in media and journalism. Digital Journalism, pp.1-10.

[3] Tjoa, S., Buttinger, C., Holzinger, K. and Kieseberg, P., 2020. Penetration testing artificial intelligence. ERCIM News, 123.

[4] Kieseberg, P., Buttinger, C., Kaltenbrunner, L., Temper, M. and Tjoa, S., 2022, July. Security considerations for the procurement and acquisition of Artificial Intelligence (AI) systems. In 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7). IEEE.