28. November 2023 – Workshop:
Sektorübergreifende Angriffsszenarien und Schutzmaßnahmen für KI-Systeme

Agenda

1. **Begrüßung**
   - Kurzes Kennenlernen (je nach Teilnehmerzahl)
   - Ziele dieses Workshops
2. **Thematische Einführung**
   - Überblick zur Normungsroadmap KI
   - Vorstellung des Bedarfs
3. **Analyse der Bedarfs-Anforderungen**
   - Diskussion Überschneidungen mit Sektoren
   - Welche Anforderungen sind sektorunabhängig, welche nicht?
   - Skizzieren von Anwendungsbereich und Gliederung
4. **Nächste Schritte**
   - Geeignete Standardisierungsform finden
   - Projekt-Initiierung

Nora Glasmeier

Technische Angriffsmöglichkeiten zur Manipulation von KI-Systemen

Umgang mit KI-Risiken

- Bei der Entwicklung und dem Einsatz von KI-Modellen müssen die **bereits bestehenden Anforderungen an die IT und an das Risikomanagement** eingehalten werden
- KI-Modelle müssen – wie auch andere Anwendungen – **in das bestehende Risikomanagement überführt werden**

Evasion Attacks

Durch eine **Manipulation von Eingabedaten** verleiten Angreifer das KI-Modell im Betrieb zu vom Entwickler **nicht vorgesehenen Ausgaben**.

Data Poisoning Attacks

Durch eine Manipulation der **Trainingsdaten** des KI-Modells erwirken Angreifer, dass dieses auf (bestimmte) Eingaben **nicht wie vom Entwickler vorgesehen reagiert**.

Privacy Attacks

Ein Angreifer **extrahiert entweder die Trainings-daten** selbst oder **Informationen über die Trainingsdaten.**

Model Stealing Attacks

Angreifer **extrahieren die Funktionalität des Modells oder greifen diese an.** Ziel ist der Diebstahl geistigen Eigentums oder die Vorbereitung anderer Angriffe.

Prompt Injection

Der Nutzer gibt dem Modell Anweisungen, durch die es sich auf **ungewünschte Weise verhält**. Im Extremfall kann dies zu **Data Poisoning** werden.

Ungewollte Abweichung der Modellgüte

Die Modellgüte ist signifikant schlechter als geplant, ohne dass dies bemerkt wird.

⚠ Angriffe zielen meist auf die **Daten in einem KI-Modell** ab. Entweder werden Daten **extrahiert** oder **manipuliert**.

Wirtschaftliche und gesellschaftliche Risiken von KI

Auch nicht-technische Risiken von KI müssen beachtet werden

- Es muss beachtet werden, dass KI-Systeme sich nicht so verhalten wie geplant. So kann es u.a. dazu führen, dass diskriminierende Tendenzen oder Falschinformationen Entscheidungsprozesse beeinflussen
- Da Cyber-Angriffe durch KI deutlich stärker werden, müssen ICT-Sicherheitsmaßnahmen ausgebaut werden

Diskriminierung und „Bias“

KI-Systeme wiederholen **diskriminierende Tendenzen oder Stereotypen** aus ihren Trainingsdaten. So werden bestimmte Gruppen benachteiligt (z.B. bei der Kreditvergabe).

Falschinformationen

KI-Systeme können **Falschinformationen** ausgeben, **Sachverhalte falsch / irreführend darstellen** oder wichtige **Informationen unterschlagen.**

Unfälle

Es ist **nicht nachvollziehbar**, wie KI-Systeme Entscheidungen treffen. So kann es dazu kommen, dass **KI-Systeme sich anders als geplant verhalten.**

KI-verstärkte Cyber-Attacken

KI kann dazu eingesetzt werden, **effizientere und wirksamere Cyber-Attacken durchzuführen.**

Verletzung der Informationssicherheit

Werden Informationen in ein externes KI-System eingegeben werden, kann **der Schutz von Informationen nicht sichergestellt werden.**

Machtkonzentration einzelner Unternehmen

Da die Entwicklung und das Trainieren von KI-Systemen **extrem ressourcenintensiv** ist, können sich **dies nur sehr einzelne Unternehmen leisten.** Diese haben praktisch keine Konkurrenz.

Grade die Tendenz der **Diskriminierung** wird im KI-Kontext sehr umfangreich besprochen.

Organisatorische Herausforderungen beim Einsatz von KI

Ausweitung des Risikomanagements

KI-Risiken müssen **in das Risikomanagement** überführt werden. Zudem müssen **vermehrt Risiko-prüfungen** durchgeführt werden, da sich KI-Systeme ständig weiterentwickelt.

Unternehmenskultur

Fail Fast und **Prototyping** muss in der **Unternehmenskultur etabliert werden.**

Vendor-Lock-In

Es gibt sehr **wenig Anbieter** von LLMs und KI-Systemen. Die Anschaff-ung ist sehr aufwändig und ein **gewünschter Anbieterwechsel sehr schwierig.**

Intervention nicht möglich

Ist es nicht möglich die Entscheidung eines KI-Systems zu ändern oder zu überschreiben, kann dies in bestimmten Fällen zu einem Risiko führen.

Dequalifizierung

Bei zunehmender Auto-matisierung geht wichtiges Wissen über Tätigkeiten verloren, so dass Mitarbei-tende ihre Aufgaben ohne das KI-System nicht mehr ausführen können.

Erklärbarkeit

Transparenz und Erklärbarkeit herzustellen ist bei KI-Systemen **extrem schwer** oder **sogar unmöglich**. Bisher sind KI-Systeme oftmals eine **Blackbox.**

Datensammlung

Zum Training von KI-Systemen sind **sehr viele Daten nötig**, weswegen auch mehr Daten gesammelt werden. Dies widerspricht dem **Grundsatz der Datenminimierung.**

Unklare Regulatorik

Der EU-AI Act wird derzeit zwischen EU-Kommission, EU-Parlament und Rat der EU verhandelt, daher **ist nicht abzusehen, was genau die Verordnung beinhalten wird.**

⚠ XXX

Bedarf 08-08: KI-spezifische Angriffsszenarien und Schutzmaßnahmen

Durch KI entsteht eine neue Risikosituation in der Finanzwirtschaft, zum einen durch die Veränderung der Intensität bestehender Risiken, aber auch durch neue Angriffsvektoren. Die veränderten Rahmenbedingungen sind in einer Normung zu berücksichtigen.

Durch den Einsatz von KI in IT-Systemen werden – unter dem Aspekt der Informationssicherheit – zusätzliche Angriffstypen und Angriffsszenarien möglich. Um das Risiko derartiger Angriffe angemessen zu reduzieren, sind diese im Rahmen von Maßnahmen zur Informationssicherheit zu beachten. Das Dokument „Sicherer, robuster und nachvollziehbarer Einsatz von KI“, welches vom BSI veröffentlicht wurde, benennt u.a. Evasion/Adversarial Attacks, Data Poisoning Attacks, Privacy Attacks, Model Stealing Attacks.

In aktuellen Normen und Standards für IT-Systeme (ohne speziellen Fokus auf die Frage, ob KI zum Einsatz kommt) wird auf diese Angriffsszenarien bzw. entsprechende Maßnahmen nicht spezifisch eingegangen. In einer Norm für KI-Systeme sollte darauf jedoch eingegangen werden.

Dieser Bedarf wird im Kontext von Finanzdienstleistungen geäußert, da die Sicherheitsanforderungen hinsichtlich Vertraulichkeit, Verfügbarkeit und Integrität (mindestens) hoch sind. Dies zeigt sich auch in bestehenden Anforderungen und Normen für allgemeine IT-Systeme durch die regulatorischen Anforderungen der Bankenaufsicht. Regelungen, die für den Einsatz von IT-Systemen (ohne Künstliche Intelligenz) sind, sind vor dem Hintergrund einer potenziell veränderten Risikosituation zu betrachten. Basierend hierauf sind zusätzliche Schutzmaßnahmen zu implementieren, die auf die konkrete Bedrohungslage abzielen.

Sektorspezifisch?

Ansätze / Zielverfolgung:

- Angriffsszenarien in „Automotive“ ähnlich wie bei Finanzwirtschaft. Später splitten und spezifizieren, aber zuerst sehr große Grundmasse von Angriffsszenarien aufgreifen → in den nächsten Jahren spezifisch einsetzen, zum jetzigen Zeitpunkt eine gemeinsame Norm anstreben.
- Meta-Standard: perspektivisch aber sektorspezifisch; wenn standardisiert bestimmte Dinge bedenken; Liste von Risiken darstellen; was KI spezifisch ist, welche Punkte müssen bedacht werden?
- Viele Querbezüge: Security-Auswirkungen auf Fairness; Robustness training für Adversarial attacks nutzen; eine Dimension herausnehmen, alle andere Dimensionen; auf Security konzentrieren.
- **Angriffsszenarien und Schutzmaßnahmen**
- **Mutwilligkeit**
  - Vergiftung von Daten
  - Rückextraktion
  - Modellmanipulation
  - (Prompt Injection)

Andere Risiken, wie normale Softwarerisiken, rauslassen. Nicht alles relevant.

→ **Fokus auf KI-Risiken!**

- Bestimmte Dinge sind KI-spezifisch, je nach Grad der Autonomie; z.B. Bias (Gefahr ist in lernendem System immer vorhanden), Transparenz wichtig
- Sektorspezifisch, Autonomiefrage besonders entscheidend; Medizin: Diagnose und Maschine entscheidet. „Degree of autonomy“ sehr wichtig!
- Cybersecurity, accuracy und autonomy gemeinsam behandeln (AI Act); nur KI-spezifische Themen

Weiterführende Fragen

- Was läuft aktuell noch? Wo einordnen? Welche ISO Dokumente gibt es? → z.B. ISO 22989 (weitere?)
- Wie robust muss Modell ausgestaltet werden? FOKUS: darlegen welche Angriffsszenarien gibt es und welche Maßnahmen müssen ergriffen werden. Ohne Schutzmaßnahmen hohe Gefahr von Bias/Fairness-Probleme → jetzt generell; dann sektorspezifisch

→ **Diskussion zu Begriffsverwirrung**

Angriffsmethoden (sind es 3…7) ?

Übergreifende Begriffe: Robustness, Fairness, Bias. Begriffe trennen, um ihre Dimension darzustellen und Verbindungen herzustellen. Die Bezüge gibt es vielfach → diese in diversen Dimensionen einholen. Begriffe auch zusammengehörend behandeln. Paper zu Angriffe (3…7?) und dies als Schwerpunkt eines Standard (DIN SPEC) machen und Ebene der Robustness und Fairness reinbringen.

Sehr klar sein in der Definition der Begriffe und Abgrenzung schaffen. Szenarien der Schwachstellen (und Angriffsvektoren) der Systeme darstellen. Zusammenfassen und ggf. neue ergänzen und dann darstellen: was ist Security/Safety und wie greift es in andere Bereiche ein?

Diskussion zum Arbeitstitel

- Schutzmaßnahmen gegen generische KI-Angriffsmethoden
- Schutzstrategien gegen generische KI-Angriffsmethoden
- **Übersicht von Angriffsmethoden auf KI-Systeme → Favorit**

Scope (Anwendungsbereich)

Dieses Dokument beschreibt Angriffsmethoden auf und Bedrohungssituationen für KI-Systeme, sowie Ziele und Auswirkungen die in diversen Sektoren auftreten können und setzt diese miteinander in Beziehung.

Abschließende Hinweise

- Was ist gemeint in jedem einzelnen Fall? Bekannte Referenzen nutzen!
- Strukturieren und in Beziehung setzen zu dem was in anderen Dokumenten vorliegt.

→ Zusammenfassung / Aufgaben:

- Titel und Scope überdenken, Ideen für Ergänzungen/Konkretisierungen sammeln → an alle
- Weiteren Termin zur Abstimmung suchen → Terminumfrage: https://nuudel.din.de/hKAdzViFcLz9skkE
- Teilnehmer für DIN SPEC abfragen → Im Rahmen der Terminumfrage
- Zeitplan: → Nächstes Treffen in Anfang 2024 → ggf. DIN-SPEC-Kick-off in Q1/2024 → Fertigstellung Q3/Q4 2024