KI-Sicherheit im Diskurs der Ergebnisse des Forschungsprojekts TAHAI (TrustAdHocAI) – Auswirkungen auf technische, organisatorische und ethische Aspekte

erstellt durch: Andreas Schmietendorf, andreas.schmietendorf@hwr-berlin.de

Hybrid durchgeführter Workshop

13. März 2025 - 13:00 bis 16:30 Uhr

Gastgeber: HTW Berlin

Campus Wilhelminenhof

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Projektleitung TAHAI:

Prof. Dr. Erik Rodner
HTW Berlin
Prof. Dr. Ralf Schnieders
HTW Berlin

Prof. Dr. Andreas Schmietendorf
HWR Berlin

Vorträge

Andreas Schmietendorf – HWR Berlin & Uni Magdeburg

Herausforderungen API-basierter KI-Dienste (Blackbox, Vorhersagbarkeit, XAI, Robustness, Fairness, …)

Ausgehend von den im Projekt TAHAI prototypisch durchgeführten KI-Implementierungen (mit Bezügen zu Bahn- und Forstwirtschaft sowie Professionsforschung) verdeutlichte der einführende Impulsvortrag einhergehende KI-Risiken (u.a. OWASP KI-Angriffsvektoren) und mögliche KI-Sicherheitsmaßnahmen. Besonderes Augenmerk galt dem Zusammenspiel zwischen KI-Sicherheit und den beim Einsatz von (Web-) APIs grundsätzlich auftretenden Sicherheits- und Compliance-Aspekten. Exemplarisch wurde dafür auf eine KI-Web-API (Image Objekterkennung) des Unternehmens OpenAI eingegangen. Diese als Black Box

zur Verfügung gestellte API gewährleistet zwar grundlegende Aspekte der IT-Sicherheit (z.B. Authentifizierung), Aussagen zur domänenspezifischen Robustheit der genutzten KI-Modelle finden sich dabei allerdings nicht.

Reiner Rumpel – Auriscon GmbH

Spezielle Angriffsvektoren / -arten von KI (Bedrohungsmodellierung)

Im Mittelpunkt des Vortrags standen die speziellen Angriffsvektoren von KI-Lösungen. Ausgehend von den klassischen Grundpfeilern Vertraulichkeit, Integrität und Verfügbarkeit erfolgte zunächst eine Klärung des Bezugsbereichs der ISO 27001 bzw. ISO 27005 (Informationssicherheits-Risikobewertung nach ISO 27001). KI-spezifische Bedrohungen sind hier allerdings kein Gegenstand der vorgenannten ISO-Norm, weshalb exemplarische Risikobetrachtungen anhand konkreter KI-Szenarien verdeutlicht wurden. Dabei wurde u.a. auf die ISO/IEC 23894:2023 (Leitlinien für das Risikomanagement bei KI-Lösungen) bzw. im Rahmen der Diskussion auf die DIN SPEC 92001-1 - Artificial Intelligence – Life Cycle Processes and Quality Requirements eingegangen.

Janek Groß, Lisa Jöckel, Michael Kläs, Pascal Gerber
Hochschule Mainz/Fraunhofer IESE

Towards Reliable AI/ML Testing
by Systematic Assessment of Test Data Quality

Im Mittelpunkt dieses Vortrags standen der Vergleich bzw. die Bewertung von KI-Modellen und die in diesem Zusammenhang benötigte Qualität eingesetzter Testdaten. Existierende Bewertungen beziehen sich aktuell eher auf deskriptive Ansätze ohne Informationen zur einhergehenden Konfidenz. Benötigt werden allerdings eher statistische Testansätze (Hypothesentest – vgl. mathematische Statistik). In Bezug auf die Testdatenqualitätsaspekte „Representativeness“, „Label Validity“ und „Self-containment“ erfolgte eine detaillierte Auseinandersetzung mit konstruktiven und analytischen Qualitätseigenschaften eingesetzter Testdaten. Dabei wurde u.a. auf methodische Ansätze wie z.B. „Confident Learning“ zum Entdecken fehlerhafter Labels eingegangen.

Rudolf Hoffmann – Hochschule für Technik und Wirtschaft Berlin

Robustheitsanalyse für Vision-Modelle

Im Beitrag geht es um bei Bild- und Image-Klassifikation ggf. auftretende Störungen und ihre Auswirkungen auf die Funktionsfähigkeit des KI-Modells. Ent-

sprechende Störungen in Bildern können z.B. durch auftretende Wolken und daraus resultierende Helligkeitsveränderungen, Positionsverschiebungen (u.a. Blickwinkel) oder auch bewusst verfälschte Labels (Label-flipping attacks) auftreten. Aussagen über die Robustheit von KI-Modellen werden z.B. als Auswahlkriterium benötigt oder auch zur Einschätzung mittels KI erzielter Ergebnisse (Konfidenz). Zur Bewertung der Robustheit von KI-Modellen wird das Framework Deepbench (Fokus auf Vision und Vision-Language Modelle) vorgeschlagen. Mit Hilfe des Frameworks lassen sich domänenspezifische Analysen zur Robustheit auf der Grundlage bewusst verfälschter Daten realisieren.

Sandro Hartenstein - HWR Berlin & Universität Magdeburg

Sicherheitsbewertung durch Angriff und Verteidigung: Ein KI-Benchmark mit präventiven Klassifikationsmethoden

Gegnerische Angriffe (adverserial attacks) auf verwendete Promts resultieren aus dem Problem der natürlich sprachlichen Verarbeitung von Daten und Befehlen, d.h. speziell ihrer nicht vorhandenen Separierung. Dem entsprechend bedarf es der Berücksichtigung potentieller Sicherheitsprobleme beim Einsatz von Large Language Modells (LLMs). Allgemein unterschieden werden direkte und indirekte Promt Injections (legitim erscheinende Cyberangriffe), Jailsbreaking (Umgehung von Zugriffsberechtigungen) und Data Leakage (unbeabsichtigter Datenabfluss). Mit dem Werkzeug promtfoo bietet sich die Möglichkeit einer Systematisierung potentiell auftretender Promt-Angriffsszenarien, die so für automatische Tests zur Robustheit von LLMs herangezogen werden können. Dafür werden eine Vielzahl kategorienspezifischer Störungen beim Benchmarking mit promtfoo berücksichtigt. Im Vortrag werden kategorienspezifische Testergebnisse (z.B. Bezüge des Promts zu chemischen und biologischen Waffen) verdeutlicht.

Dank

Allen Vortragenden und Gästen (insgesamt waren wir 15 Teilnehmer) sei für ihre Mitwirkung und vor allem die konstruktiv geführten Diskussionen gedankt. Ein besonderer Dank gilt dem Gastgeber des Workshops Herrn Prof. Dr. Erik Rodner für die ausgezeichneten Rahmenbedingungen bzw. die Bereitstellung des köstlichen Kaffees und Kuchens!