Vertrauenswürdigkeit von KI-Lösungen (Implikationen im Data Science und Software-Engineering)

detaillierter Bericht zum hybrid durchgeführten ESAPI-Workshop 2022

Andreas Schmietendorf, Jens Heidrich
Email: andreas.schmietendorf@hwr-berlin.de, jens.heidrich@iese.fraunhofer.de

1. Motivation zum Workshop

Das Vertrauen in Anwendungen der künstlichen Intelligenz ist von multidimensionalen Aspekten abhängig. Die „Ethics Guidelines for Trustworthy AI“ der Europäischen Kommission definieren verschiedene Prinzipien und Handlungsempfehlungen, wie das Abwenden von Schaden, Fairness oder transparente Prozesse, als Grundlage für Vertrauenswürdigkeit. Eine ausschließliche Berücksichtigung der technischen Eigenschaften entwickelter Lösungen, die sich z.B. an der ISO 25000 (SQuaRE - System and Software Quality Requirements and Evaluation) orientiert, ist zwar sinnvoll, reicht aber zur Gewährleistung vertrauenswürdiger KI-Lösungen nicht aus. Die VDE-Anwendungsregel VDE-AR-E 2842-61 des DKE-Arbeitskreises 801.0.8 bricht Vertrauenswürdigkeit in einzelne Qualitätsaspekte herunter (vgl. [Johner 2022]):

> “Trustworthiness […] combines several aspects of trustworthiness in a quite generic way: for every product the set of aspects can be suitably selected and remains unchanged throughout the project. Aspects of trustworthiness include but are not limited to system safety, functional safety, safety of use, security, usability, ethical and legal compliance, reliability, availability, maintainability, and (intended) functionality.“
>
> (Originalquelle: *VDE-AR-E 2842-61-1 Kapitel 3.1.43*)

Mit Hilfe von KI-Lösungen gewonnene Texte, Klassifizierungen, Prognosen oder auch Bild-, Audio- und Videoanalysen implizieren Bedürfnisse hinsichtlich der Erklär-, Interpretier- und Reproduzierbarkeit. Dabei geht es nicht zuletzt um die Vermeidung diskriminierender Ergebnisse durch den Einsatz von KI-Algorithmen. Die Reproduzierbarkeit erzielter Analyseergebnisse wird durch das BSI als direkte Vorrausetzung für die Verbreitung vertrauenswürdiger KI-Ansätzen genannt (vgl. [BSI 2022] S. 3):

> „Furthermore, reproducibility is a requirement for establishing causality for the interpretation of model results and building of trust towards the overwhelming expansion of AI systems applications.“ (Quelle: [BSI 2022])

Unter Berücksichtigung der aufgezeigten Komplexität des Begriffs der Vertrauenswürdigkeit im KI-Diskurs bedarf es dennoch einfach zu handhabender Prinzipien und Methoden, die eine Auseinandersetzung mit sinnfälligen KI-Lösungen nicht von vornherein obsolet machen. Ziel des Workshops war es, sich mit sowohl mit praxis- als auch forschungsorientierten Aspekten auseinanderzusetzen, wie bespielhaft den folgenden:

- Vielfältige praxisrelevante KI-Anwendungsszenarien:
    - Sentiment-Analysen für ein besseres Kundenverständnis,
    - Bewältigung massenhafter Problem-Tickets (Klassifikation),
    - Bild- und Videoverarbeitung zur Gefahrenerkennung,
    - Mustererkennung zur Identifikation von Krebszellen,
    - …
- Forschungsorientierte KI-Fragen
    - Vertrauen in KI-Algorithmen aus der „Steckdose“,
    - Test- und Erklärbarkeit von KI-Ergebnissen,
    - Messbarkeit qualitativer Eigenschaften eingesetzter KI-Algorithmen,
    - KI als Unterstützung im Software-Engineering bzw. Reengineering,
    - …

2. Beiträge des Workshops

Durch die Initiatoren des Workshops erfolgte zunächst eine motivierende Einführung in das Themengebiet.

- Als Gastgeber der Veranstaltung ging *Dr. Jens Heidrich* (Fraunhofer IESE - Division Manager Smart Digital Solutions) auf die vielfältigen Bedürfnisse verlässlicher und sicherer KI-Systeme aus Sicht von Industrie- und Forschungsprojekten ein.
- In seiner Rolle als Sprecher der GI-Fachgruppe "Measurement & Data Science" (FG 2.1.10) unterstrich *Dr. Andreas Jedlitschka* (Fraunhofer IESE - Department Head Data Science) den Bedarf an nachhaltig einsetzbaren KI-Methoden im Bereich des System- und Software-Engineerings.
- Eröffnet wurde der Workshop schließlich von *Prof. Dr. Andreas Schmietendorf* (Sprecher der ESAPI-Initiative). Im Mittelunkt seiner Ausführungen standen der Einsatz problemadäquater (KI-) Algorithmen, der Bedarf an agil durchgeführten KI-Experimenten bzw. Tests, die notwendige Akzeptanz lernender Projektorganisationen, aber auch die Gewährleistung von Sicherheit, Vertrauen und Compliance zu geltenden Standards, Regeln und Gesetzen.

Im Vorfeld des Workshops konnten die folgenden eingeladenen Gastredner für die Vormittagssession gewonnen werden. Bei der zusammenfassenden Darstellung der Vortragsinhalte handelt es sich um die Interpretation der Autoren dieses Berichts:

Dr. Gaby Gurczik
Referentin für Grundsätze KI und Datenökonomie beim BMDV

Titel: KI-Innovationen als Standortchance für Deutschland und Europa

Im Kern beschäftigte sich der Vortrag mit den benötigten (politischen) Rahmenbedingungen für innovative KI-Anwendungen im Mobilitätsbereich. In diesem Zusammenhang wurde u.a. auf rechtliche Aspekte, wie den EU AI Act, Gefahren einer Überregulierung, den Bedarf offener und qualitativ hochwertiger Daten-Ökosysteme sowie benötigte sinnstiftende aber dennoch erklärbare Anwendungsszenarien intelligenter Mobilitätsdienste, eingegangen.

Dr. Rasmus Adler
Leiter des Programms Autonome Systeme am Fraunhofer IESE

Titel: Das Spaltmaß für KI-Systeme - Wie sieht es aus und was sind akzeptable Grenzwerte?

„KI aus Deutschland soll zum Gütesiegel werden“ - Die strategische Fokussierung auf Qualität, Regulatorik und Normierung passt zur historischen Verortung des deutschen Industriestandorts. Die dafür benötigte konsistente Definition eines KI-Systems existiert aktuell jedoch nicht, was mit vielfältigen Widersprüchen zwischen EU AI Act und der ISO/IEC 22989 im Vortrag belegt wurde. Darüber hinaus wurde der Bedarf messbarer Qualitätsaspekte über die verschiedenen Systemebenen hinweg verdeutlicht.

Prof. Dr. Katharina Zweig
Leiterin Algorithm Accountability Lab TU Kaiserslautern

Titel: Kann man mit Surrogatansätzen KI-Entscheidungen erklären?

Der Bedarf erklärbarer KI-Ergebnisse bezieht sich u.a. auf die Plausibilisierung von Kausalbeziehungen, die Erhöhung des Informationsgehalts, die Gewährleistung fairer (nichtdiskriminierender) Entscheidungen oder auch die Festlegung klarer Verantwortlichkeiten. Mit Hilfe eines trivialen KI-Szenarios zur Kreditwürdigkeit (entsprechend angelernte Entscheidungsbäume) wurde die Untauglichkeit des Einsatzes eines ebenfalls entscheidungsbaumbasierten Surrogatmodellansatzes verdeutlicht, da es ungerechtfertigte Entscheidungen verschleiern kann.

Die Nachmittagssession war dann der gemeinsamen Diskussion vorbehalten, wobei verschiedene Themen durch korrespondierende Impulsvorträge eingeführt wurden:

Sandro Hartenstein

Titel: Vertrauenswürdige KI-Web-API Spezifikationen

Inwieweit lassen sich mit der OpenAPI-Spezifikation (ehem. Swagger) Attribute vertrauenswürdiger KI-Web-APIs abbilden. Mit Hilfe einer GAP-Analyse wurden nicht abgedeckte Attribute (z.B. Safety, Transparenz, Diskriminierungsfreiheit) verdeutlich und grundlegende Ansätze zur Integration dieser Angaben aufgezeigt.

Julius Schinschke

Titel: LoRaWAN Netzabdeckungsmessungen im Kontext der Vertrauenswürdigkeit

Eingesetzte Sensorik im Diskurs von IoT-Lösungen (Internet of Things) können gewonnene Daten mit Hilfe des Long Range Wide Area Network energieeffizient übertragen. Für nachhaltig betriebene Anwendungsszenarien bedarf es allerdings der Kenntnis geografisch verorteter Empfangsfeldstärken konkreter LoRaWAN-Gateways.

Lukas Scholz

Titel: Explainable AI – Analyse und Realisierung zur Erklärbarkeit von Computer-Vision-Modellen

Zunächst wurde auf verschiedene Ansätze zur Erklärbarkeit eingegangen (Deconvolutional Networks, Guided Backpropagation und Class Activation Maps – kurz CAM). Im Weiteren setzte sich der Beitrag mit der Konzeption (Date, Training, Visualisierung) und prototypischen Implementierung eines erklärbaren AI-Systems (CAM-basiert) auseinander.

Daniel Krohmer

Titel: Software Marketplaces for Extensible Web Apps

Marktplätze für Software implizieren hohe Risiken hinsichtlich Qualitätsverletzungen (z.B. Cross-side Scripting), die zumeist viele Konsumenten betreffen. Dieser Aspekt wurde mit historischen Daten zu PlugIns für das CMS WordPress verdeutlicht. Die Idee ist es, derartige Probleme (Muster) für das Training von „Machine Learning“-basierter Schwachstellenanalysatoren heranzuziehen.

Dr. Michael Kläs et al.

Titel: An Assurance Case Pattern to Argue Quantitative Safety Targets for AI Components …

Der Beitrag fokussierte auf die Verwendung risikogetriebener Akzeptanzkriterien zur Strukturierung von Assurance Cases bei sicherheitskritischen KI-Komponenten. Mit Hilfe integrierter Messansätze innerhalb des Designs bzw. der genutzten Laufzeitumgebungen soll das Erreichen festgelegter Sicherheitsziele nachgewiesen bzw. argumentiert werden.

3. Weitere Informationen

Viele der hier besprochenen Beiträge können auf der korrespondierenden Webseite zum Workshop mittels des folgenden QR-Codes heruntergeladen werden:

Darüber hinaus sei auf den im Rahmen der GI-Jahrestagung INFORMATIK 2023 für den 27. September 2023 an der HTW Berlin geplanten Workshop „Young Scientists and early-stage research in Data Science Workshop 2023 (YSDS-23)“ verwiesen (vgl. folgender QR Code), noch ist es Zeit sich mit einem Beitrag zu beteiligen!

Obwohl die Möglichkeiten, aber auch Limitierungen KI-basierter Chat-BOTS für die Informatik-Community nicht überraschend waren, zeigt der Hype um das prototypische Angebot von chatGPT (vgl. https://chat.openai.com/auth/login) die Dynamik und das zunehmende öffentliche Interesse im Bereich der KI-Themen. Grund genug auch im Jahr 2023 eine Neuauflage des ESAPI-Workshops zu planen. Weiterführende Informationen werden zeitnah unter der URL: https://blog.hwr-berlin.de/schmietendorf/ bzw. der Webseite der GI-Fachgruppe "Measurement & Data Science" (https://fg-data-science.gi.de) bereitgestellt.

4. Quellenverzeichnis

[Johner 2022] Johner, C.: Weshalb die VDE-AR-E 2842-61 (vertrauenswürdige KI-Systeme) nicht nur die Entwicklung betrifft, 13. April 2021, https://www.johner-institut.de/blog/systems-engineering/ki-systeme/#section_scroll3, letzter Zugriff: Februar 2023

[BSI 2022] Deep Learning Reproducibility and Explainable AI (XAI) Results of BSI's project research, Federal Office for Information Security 2022, https://www.bsi.bund.de, letzter Zugriff 13. September 2022

5. Partner

Fraunhofer-Institut für Experimentelles Software Engineering (Gastgeber)
https://www.iese.fraunhofer.de

Fachgruppe Measurement & Data Science
https://fg-data-science.gi.de/

Central Europe Computer Measurement Group (Sponsoring)
https://cecmg.de

Arbeitskreis Software-Qualität und -Fortbildung e.V. (ASQF)
www.asqf.de

SIGS DATACOM GmbH (Medienpartner)
https://www.sigs-datacom.de

Shaker Verlag GmbH Düren (Medienpartner)
https://www.shaker.de

Dank

Unser Dank gilt den Referenten, Teilnehmern und den Partnern und Sponsoren (Fraunhofer IESE Kaiserslautern, HWR Berlin und der ceCMG e.V.), die eine solche Veranstaltung ermöglicht haben. Ein besonderer Dank gilt Herrn Dr. Jens Heidrich für das Management der ausgezeichneten Rahmenbedingungen vor Ort!