

KI-Sicherheit im Diskurs domänenspezifischer Anwendungsfelder

Sandro Hartenstein
sandro.hartenstein@hwr-berlin.de

Herausforderungen Low-Code orientierter KI-Ansätze
Fraunhofer IESE Kaiserlautern
12.11.2024

Agenda



- Motivation
- Umfrage
- Diskussionsrunden – KI-Sicherheit (je ca 10 Minuten)
 - Runde 1: Sprachverarbeitung mittels LLMs in der Mediationsanalyse
 - Runde 2: Totholzerkennung via Drohnen
 - Runde 3: Vandalismus-Erkennung und visuelle Echtzeitklassifikationsalgorithmen
- Zusammenfassung und Abschluss

Motivation



Quelle: <https://chatgpt.com/share/672dd785-d4bc-800d-a98f-eb983341071f>



Der zunehmende Einsatz von KI-Systemen in sicherheitskritischen und sensiblen Anwendungsbereichen stellt uns vor neue Herausforderungen.

Wir betrachten drei konkrete Anwendungsfälle:

- Die Analyse vertraulicher Mediationsgespräche mittels Large Language Models
- Die automatisierte Erkennung von Totholz durch Drohnensysteme im Forst
- Die KI-gestützte Videoanalyse zur Erkennung von Vandalismus

Aktuelle Relevanz

Diese Anwendungsfälle repräsentieren unterschiedliche Facetten moderner KI-Systeme:

- Verarbeitung sensibler personenbezogener Daten
- Autonome Entscheidungsfindung in der realen Umgebung
- Automatisierte Überwachung und Analyse im öffentlichen Raum

Forschungsbedarf

Die Entwicklung geeigneter Sicherheitsstandards für diese Systeme erfordert

- Technische Robustheit und Zuverlässigkeit
- Rechtliche und datenschutzrelevante Konformität
- Ethische Vertretbarkeit und gesellschaftliche Akzeptanz

Bedeutung des interdisziplinären Austauschs

Ihre unterschiedlichen Expertisen sind entscheidend für:

- Identifikation kritischer Sicherheitsanforderungen
- Entwicklung praxistauglicher Lösungsansätze
- Bewertung möglicher Zielkonflikte

Ziele der heutigen Diskussionsrunde sind

- Verknüpfung verschiedener Fachperspektiven
- Aufzeigen von Ansätzen für sichere KI-Systeme

Umfrage 5min

<https://www.menti.com/al77parxiw1v>





Diskussionsrunden

Der Teamsstream soll aufgezeichnet, später transkribiert und für Forschungszwecke ausgewertet werden.

Ziel der Analyse ist es, die Differenzen der anwendungsbezogenen Anforderungen an KI-Sicherheit aufzuzeigen.

Runde 1: Mediationsanalyse und LLMs



Einordnung:

- KI-gestützte Analyse von Mediationsgesprächen

Technischer Kontext:

- Large Language Models analysieren Gesprächsprotokolle

Aktuelle Herausforderungen:

- Vertraulichkeit sensibler Gesprächsinhalte
- Zuverlässigkeit der KI-Analyse
- Balance zwischen Analysetiefe und Datenschutz

Diskussionsziel:

- Ansätze für vertrauenswürdige Analysen

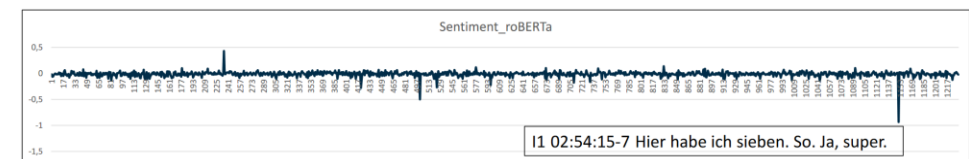
Was sind die Streitpunkte

Basierend auf den gegebenen Kontexten lassen sich folgende Streitpunkte identifizieren:

- 1. Betreuung vs. Freiheit in der Entscheidung, wer die Betreuung übernimmt** (Context 0): Hier geht es um die Präferenz, wer die Betreuung einer bestimmten Person übernehmen soll, und die Abwägung dieser Entscheidung gegenüber dem Wunsch oder der Freiheit, diese Betreuung selbst zu übernehmen.
- 2. Klärung finanzieller Aspekte** (Context 1 & 2): In diesen Kontexten wird ein Streitpunkt über die Klärung oder Abklärung finanzieller Details angesprochen, speziell um den Anruf bei einer bestimmten Person namens Roman, um den genauen Betrag zu erfragen, der für eine bestimmte Sache oder Leistung relevant ist.
- 3. Bedeutung von Erklärungen für das Verständnis von Reaktionen** (Context 3): Der Streitpunkt hier dreht sich um die Notwendigkeit, Erklärungen für bestimmte Reaktionen oder Verhaltensweisen zu liefern. Eine Partei fühlt sich missverstanden ohne die Bereitstellung von Kontext oder Erklärung für ihr Verhalten, während die andere Partei der Meinung ist, keine Erklärungen zu benötigen.

Diese Streitpunkte reflektieren unterschiedliche Bedürfnisse und Perspektiven der Beteiligten, die während der Mediation adressiert und möglichst aufgelöst werden sollen.

Retrieval Augmented Generation Fall5



Sentiments im Zeitverlauf Fall5 S3

Runde 2: Totholzerkennung via Drohnen



Einordnung:

- Automatisierte Waldinspektion mittels Drohnen

Technischer Ansatz:

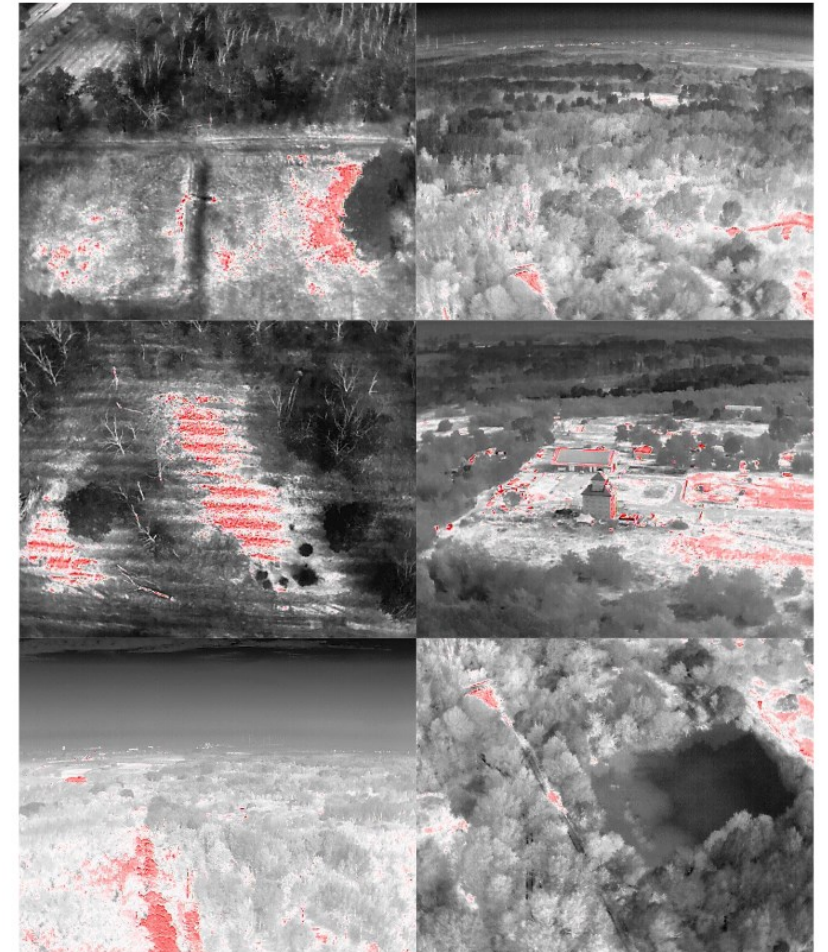
- Visuelle KI-Klassifikation von Luftbildern (Wärme- und RGB Bilder)

Zentrale Aspekte:

- Präzision der Erkennung
- Sicherheit im Flugbetrieb
- Umgang mit Umwelteinflüssen

Diskussionsziel:

- Ansätze für robuste visuelle KI-Systeme



Quelle: <https://blog.hwr-berlin.de/schmietendorf/wp-content/uploads/2024/07/rodner-tahai-jul2024.pdf>

Runde 3: Vandalismus-Erkennung und visuelle Klassifikationsalgorithmen

Einordnung:

- KI-gestützte Videoüberwachung öffentlicher Räume

Technische Ansatz:

- Visuelle KI-Algorithmen zur Echtzeiterkennung von Objekten und Veränderungen

Zentrale Aspekte :

- Echtzeitanalyse von Videodaten
- Balance zwischen Sicherheit und Privatsphäre
- Vermeidung von Fehlalarmen

Diskussionsziel:

- Ansätze für sichere Überwachung mit KI



Quelle: <https://universe.roboflow.com/nimra-sardar-n5hc6/vandalism-detection-2/images/qzbGuR9MO6RIV85OHovj>

Zusammenfassung und Abschluss



Vielen Dank!

Sandro Hartenstein
sandro.hartenstein@hwr-berlin.de

Herausforderungen Low-Code orientierter KI-Ansätze
Fraunhofer IESE Kaiserlautern
21.11.2023