2.4.3 Vertrauenswürdige KI und Sicherheit

Themengruppe 2

Exzellenz & Vernetzung - Wissenschaftliches Kompetenzfeld Künstliche Intelligenz / Data Science

Projektbeschreibung

Die KI und insbesondere das maschinelle Lernen (ML) sind wegen der Komplexität der Algorithmen, der sehr langen und ineinander verzahnten Prozesse innerhalb eines Lernprozesses und der Abhängigkeit auch von zeitlich veränderlichen Daten für die menschlichen Anwenderinnen und Anwender nicht leicht zu verstehen. Es geht darum, Zusicherungen zur Robustheit und Güte gelernter Modelle machen zu können. Welche Anforderungen haben die verschiedenen Anwendungen von BAuA und IFR? Das Kompetenzzentrum ML2R untersucht Zertifzierungsansätze und entwickelt Testmethoden. Hier bietet sich eine Zusammenarbeit mit BAuA und dem IFR an.
ML2R und die Informatik der TU Dortmund arbeiten auch an Erklärungen gelernter Modelle, der SFB 876 entwickelt Zusammenfassungen großer Datenströme für den raschen Überblick.
Umgekehrt kann ML auch helfen, Probleme für die Sicherheit von Prozessen oder Dingen zu erkennen. Dies kann beispielsweise durch Die Analyse von Texten geschehen, in denen auf Mängel von Objekten hingewiesen wird.

Projektziele

- Ansätze zur Zertifizierung zusammen bringen
- Sammlung von gewünschten Zusicherungen
- Erklärungen von Modellen, Datenzusammenfasung
- Erkennung von Sicherheitsproblemen durch ML

Meilensteine/Zeitplan

Ein Bericht über die Anforderungen und vorhandenen Ansätze zur Zertifizierung und Erklärung von Modellen des maschinellen Lernens könnte schon im Herbst 2021 vorliegen.

Mitwirkende

Ansprechpartner*in

Dr. Stefan Michaelis, TU Dortmund

Partner*innen

- Prof. Dr. Lars Adolph, BAuA
- Martin Goetzke, IFR
- Prof. Dr. Katharina Morik, TU Dortmund
- Prof. Dr. Emmanuel Müller, TU Dortmund
- Prof. Dr. Erich Schubert, TU Dortmund
- MPI CyberSecurity