TrustAI

A Platform for Interactive and Trustworthy Artificial Intelligence

Programm / Ausschreibung	Kooperationsstrukturen, Kooperationsstrukturen, FH - Forschung für die Wirtschaft (COIN-Aufbau) Ausschreibung 2022	**Status**	laufend
Projektstart	01.05.2023	**Projektende**	30.04.2027
Zeitraum	2023 - 2027	**Projektlaufzeit**	48 Monate
Keywords	Interactive Machine Learning; Explainable Machine Learning; Human-Machine Collaboration; Trustworthy AI		

Projektbeschreibung

Künstliche Intelligenz (KI/AI) basierend auf maschinellem Lernen transformiert derzeit praktisch alle Sektoren in Industrie und Wirtschaft. Laut Gartner, Inc. wird der weltweite Umsatz mit KI 2022 voraussichtlich 62,5 Mrd. Dollar betragen, was einem Anstieg von 21,3 % gegenüber 2021 entspricht. Speziell KMUs haben heute das Problem, dass Sie existierende AI-Lösungen nur unzureichend nutzen und so den mit AI generierbaren Mehrwert nicht heben können. Die Hauptgründe dafür sind:

- hochspezifische Analyseanforderungen von KMUs für die es keine gebrauchsfertigen Lösungen gibt;
- mangelnde Adaptionsfähigkeit existierender AI-Services an sich dynamisch ändernde Anforderungen;
- intransparente Entscheidungsfindung existierender AI-Services und damit mangelndes Vertrauen;
- unzureichende Menge an annotierten Daten für das Training der AI;
- zu wenig AI-bezogenes Knowhow in vielen KMUs für eine firmeninterne Realisierung.

Diese Faktoren hemmen den Einsatz von AI in KMUs sowie die Entwicklung von vertrauenswürdigen und resilienten KI-Lösungen.

Ziel des Projekts ist es, interaktiv trainierbare und dynamisch anpassbare AI-Services für die hochspezifischen Anforderungen von KMUs zu entwickeln, die durchgängig transparent und daher vertrauenswürdig sind. Unsere zentrale Innovation dabei ist ein neues interaktives Trainingsparadigma, in dem Mensch und AI in einem wechselseitigen Dialog stehen und durch gegenseitige Erklärungen ein gemeinsames Verständnis von Problemen entwickeln. So werden zwei Ziele zugleich erreicht: eine flexible vom Menschen gesteuerte Anpassung der AI an hochspezifische Anforderungen und eine inhärent sich selbst erklärende AI die vertrauenswürdig agiert. Dieser Lernansatz ist rein datengetriebenen Lernmethoden (Supervised Learning) und klassischen Human-in-the-Loop Ansätzen (Active Learning) überlegen, da er zum einen das Lernen auf Basis kleiner annotierter Datensätze ermöglicht und zum anderen kontinuierlich die Transparenz der KI sicherstellt.

Angestrebte Ergebnisse umfassen:

- Proof-of-Concept des vorgeschlagenen interaktiven und transparenten KI-Ansatzes.
- Eine Service-Plattform inkl. Interfaces, die unseren AI-Ansatz für KMUs niederschwellig zur Verfügung stellt. Die Plattform

soll sowohl den initialen Lernprozess als auch die laufende Adaptierung, Verifikation und Konsolidierung der AI abdecken (MLOps).
- Evaluierung des Nutzens unseres Ansatzes und der Plattform in spezifischen heterogenen Fallstudien in den Domänen Gesundheit, Landwirtschaft und kulturelles Erbe.
Das Projekt wird am neuen Center for Artificial Intelligence der Fachhochschule St. Pölten umgesetzt und unterstützt den Kompetenzaufbau in den zwei Forschungsschwerpunkten Human-Centered AI und Trustworthy AI. Die im Projekt entwickelten Lösungen sollen KMUs helfen, effizient hochspezifische und gleichzeitig vertrauenswürdige AI-Lösungen zu entwickeln, um ihre Wettbewerbsfähigkeit zu erhöhen.

Abstract

Artificial intelligence (AI) based on machine learning is currently transforming virtually all sectors in industry and business. According to Gartner, Inc., global AI revenue is expected to reach $62.5 billion in 2022, up 21.3% from 2021. Today, SMEs in particular face the problem of underutilizing existing AI solutions and thus failing to leverage the value that can be generated with AI. The main reasons for this are:
- highly specific analysis requirements of SMEs for which there are no ready-to-use AI solutions;
- lack of adaptability of existing AI services to dynamically changing requirements;
- non-transparent decision making of existing AI and thus lack of trust;
- insufficient amount of annotated data for AI training; and
- too little AI-related know-how in many SMEs for in-house implementation.
These factors inhibit the use of AI in SMEs and the development of trustworthy and resilient AI solutions.
The goal of this project is to develop interactively trainable and dynamically adaptable AI services for the highly specific needs of SMEs that are end-to-end transparent and therefore trustworthy. Our central innovation is a new interactive training paradigm in which humans and AI engage in a two-way dialogue and develop a common understanding of problems through mutual explanations. This achieves two goals at once: flexible human-driven adaptation of AI to highly specific requirements and an inherently self-explanatory AI that acts in a trustworthy manner. This learning approach is far superior to purely data-driven learning methods (supervised learning) and classical human-in-the-loop approaches (active learning) because, on the one hand, it enables learning based on small datasets and, on the other hand, it ensures ongoing transparency and trustworthiness of the AI.
Intended outcomes include:
- Proof-of-concept of the proposed interactive and transparent AI approach.
- A service platform including interfaces that make our AI approach easily available to SMEs in a low-threshold manner. The platform should cover the initial learning process as well as the ongoing adaptation, verification, and consolidation of the AI (MLOps).
- Evaluation of the utility of our approach and the platform in specific heterogeneous case studies in the domains: health, agriculture, and cultural heritage.
The project will be implemented at the new Center for Artificial Intelligence at St. Pölten University of Applied Sciences and will support competence building in the two main research areas Human-Centered AI and Trustworthy AI. The solutions developed in the project will help SMEs to efficiently develop highly specific and at the same time trustworthy AI solutions to increase their competitiveness.

Projektpartner

- Hochschule für Angewandte Wissenschaften St. Pölten Forschungs GmbH