ABRRA

Algorithmic Bias Risk Radar

Programm / Ausschreibung	IWI 24/26, IWI 24/26, Basisprogramm Ausschreibung 2024	**Status**	abgeschlossen
Projektstart	01.06.2024	**Projektende**	30.11.2025
Zeitraum	2024 - 2025	**Projektlaufzeit**	18 Monate
Keywords			

Projektbeschreibung

Das Projekt Algorithmic Bias Risk Radar (ABRRA) des Startups leiwand AI in Zusammenarbeit mit der TU Wien mit einer geplanten Laufzeit von 24 Monaten verfolgt eine hochinnovative Zielsetzung im stark wachsenden Markt AI-Vertrauens-, Risiko- und Sicherheitsmanagement (AI TRiSM): Basierend auf der erstmaligen Entwicklung einer hochwertigen Experten-Datenbank zu Diskriminierungs- und Bias-Risiken von AI-Systemen in englischer Sprache in Form von sogenannten AI Incidents wird unter Anwendung von Knowledge Graphs, statistischen Verfahren und aktuellen Machine Learning-Methoden (u.a. LLM) ein Prototyp für ein neuartiges Pre-Assessment von Diskriminierungs- und Bias-Risiken in AI-Anwendungen entwickelt.
Beim Pre-Assessment, das durch den Algorithmic Bias Risk Radar möglich wird, handelt es sich um ein Softwareprodukt, das Expert:innen erstmals die Durchführung von Ähnlichkeits- und Musteranalysen in Bezug auf Bias-Risiken von AI-Systemen ermöglichen wird und das weltweit noch nicht am Markt erhältlich ist. Wesentliche Entwicklungsrisiken liegen in der Erzielung einer hohen Datenqualität, Umgang mit unvollständigen Daten und Sicherstellung einer hohen Ergebnisqualität unter Vermeidung von Bias und Halluzinationen.
Der Algorithmic Bias Risk Radar ist ein wesentlicher Beitrag dazu, dass künftig nachteilige Wirkungen von AI-Anwendungen im Zuge der Entwicklung, Beschaffung und Zertifizierung früher erkannt, analysiert und mitigiert und dass die im neuen EU AI Act verlangten Fundamental Rights Impact Assessments bei High-Risk-Anwendungen z.B. in Human Ressources, Gesundheit, Finanzen und öffentlicher Verwaltung zielgerichtet durchgeführt werden können.
Das Projekt wird in enger Kooperation zwischen Sozialwissenschafter:innen, welche Wissen über Bias und Diskriminierung in Anwendungsfeldern einbringen, Data Scientists und ML-Expert:innen durchgeführt.

Endberichtkurzfassung

Das Projekt Algorithmic Bias Risk Radar (ABRRA) des Startups leiwand AI in Zusammenarbeit mit der TU Wien entwickelte eine neuartige AI Bias Incident Datenbank nach wissenschaftlichen Standards als Basis eines KI-Bias-Früherkennungs-Tools. Sie ermöglicht ein neuartiges Pre-Assessment von Diskriminierungs- und Bias-Risiken durch KI-Systeme in unterschiedlichen Anwendungsbereichen. Die Datenbank entspricht höchsten Anforderungen an Datenqualität, Labeler-Agreement und sozialwissenschaftlichen Standards.

Projektkoordinator

- leiwand AI gmbh

Projektpartner

- Technische Universität Wien