CIMPLE

Countering Creative Information Manipulation with Explainable AI

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft - Vorbereitung Horizon Europe	**Status**	abgeschlossen
Projektstart	01.04.2021	**Projektende**	30.06.2024
Zeitraum	2021 - 2024	**Projektlaufzeit**	39 Monate
Keywords	Explainable AI, Deep Learning, Knowledge Graphs, Information Manipulation, Visualization		

Projektbeschreibung

Die Erklärbarkeit ist von großer Bedeutung auf dem Weg zu einer vertrauenswürdigen und ethischen KI, doch sie steckt derzeit noch in den Kinderschuhen. Relevante Bemühungen konzentrieren sich meist auf erhöhte Transparenz der KI-Modellgestaltung sowie auf die Interpretationen der resultierenden Entscheidungen. Der Verständlichkeit solcher Erklärungen und ihrer Eignung für bestimmte Szenarien und Anwender wurde bisher nur wenig Aufmerksamkeit geschenkt. CIMPLE hat die drastische und interdisziplinäre Weiterentwicklung von XAI-Methoden zum Ziel, um diese rekonfigurierbar und personalisierbar zu machen.

Wissensgrafiken bieten signifikantes Potenzial, um KI-Modelle besser zu strukturieren. Indem sie Kontext und Anwendungsdomäne in granularer Weise erfassen, bieten solche Graphen eine dringend benötigte semantische Schicht, die derzeit bei maschinellen Lernverfahren fehlt.

Menschliche Faktoren sind wichtige Determinanten für den Erfolg von KI-Modellen. In bestimmten Anwendungsbereichen, wie z.B. bei der Erkennung manipulierter Information, reichen die vorhandenen technischen XAI-Erklärungsmethoden nicht aus, da die Komplexität der Domäne und eine Vielzahl sozialer und psychologischer Faktoren das Vertrauen der Benutzer in abgeleitete Erklärungen negativ beeinflussen können. In der Vergangenheit hat die Forschung gezeigt, dass es wenig effektiv ist Benutzern wahr/falsch Entscheidungen zu präsentieren, insbesondere wenn ein Black-Box-Algorithmus verwendet wird. Zu diesem Zweck will CIMPLE mit innovativen sozialen und wissensbasierten „kreativen“ KI-Erklärungen experimentieren. Diese werden im Bereich der Erkennung und Verfolgung manipulierter Informationen getestet, wobei sowohl technische als auch soziale und psychologische Anforderungen berücksichtigt werden.

Abstract

Explainability is of significant importance in the move towards trusted, responsible and ethical AI, yet remains in infancy. Most relevant efforts focus on the increased transparency of AI model design and training data, and on statistics-based interpretations of resulting decisions. The understandability of such explanations and their suitability to particular users and application domains received very little attention so far. Hence there is a need for an interdisciplinary and drastic evolution

in XAI methods, to design more understandable, reconfigurable and personalisable explanations.

Knowledge Graphs offer significant potential to better structure the core of AI models, and to use semantic representations when producing explanations for their decisions. By capturing the context and application domain in a granular manner, such graphs offer a much needed semantic layer that is currently missing from typical brute-force machine learning approaches.

Human factors are key determinants of the success of relevant AI models. In some contexts, such as misinformation detection, existing XAI technical explainability methods do not suffice as the complexity of the domain and the variety of relevant social and psychological factors can heavily influence users' trust in derived explanations. Past research has shown that presenting users with true / false credibility decisions is inadequate and ineffective, particularly when a black-box algorithm is used. To this end, CIMPLE aims to experiment with innovative social and knowledge-driven AI explanations, and to use computational creativity techniques to generate powerful, engaging, and easily and quickly understandable explanations of rather complex AI decisions and behaviour. These explanations will be tested in the domain of detection and tracking of manipulated information, taking into account social, psychological and technical explainability needs and requirements.

Projektpartner

- webLyzard technology gmbh