Karlsruher Institut für Technologie

Institut für Arbeitswissenschaft und Betriebsorganisation

Engler-Bunte-Ring 4
76131 Karlsruhe

Bachelorarbeit / Masterarbeit

Explainable Artificial Intelligence: Effekte personalisierter Erklärungen auf Vertrauen in und Verstehen von Systemen

Durch die voranschreitenden Entwicklungen im Maschinellen Lernen ist es möglich, maschinelle Ausgaben weitgehend automatisiert und rein datenbasiert zu erzeugen. Die Ausgaben solcher Systeme sind – meist sogar für Fachkräfte – nicht interpretierbar. KI-Systeme werden somit zu einer Black-Box und erschweren dem Menschen, einen Einblick in ihre Funktionsweise zu bekommen und die Ausgaben inhaltlich zu verstehen.

Erklärbare Künstliche Intelligenz (XAI) umfasst verschiedene Prozesse und Methoden, die es Menschen ermöglichen sollen, die von maschinellen Lernalgorithmen erzeugten Ergebnisse und Ausgaben zu verstehen. XAI-Ansätze stellen somit eine Möglichkeit dar, Verständnis und Nachvollziehbarkeit auf Seite der Anwendenden zu ermöglichen, mit dem Ziel, Vertrauen in die Systeme und dadurch eine langfristige Nutzung zu erreichen.

Im Rahmen des BMBF-geförderten Forschungsprojekts KARL (*Künstliche Intelligenz für Arbeit und Lernen in der Region Karlsruhe, hier geht's zum Projekt*) wird in Zusammenarbeit mit dem Fraunhofer IOSB in einer experimentalpsychologischen Studie untersucht, wie sich Möglichkeiten der Personalisierung von Erklärungen auf Menschen auswirken. Im Rahmen der Studie kommt eine am IOSB entwickelte Anwendung zum Einsatz, in der auf einem künstlichen Datensatz basierende kontrafaktische Erklärungen durch Gewichtungen und Ein-/Ausschluss von Faktoren personalisiert werden können.

Ziel der Studie ist es herauszufinden, ob personalisierte Erklärungen zu einem besseren lokalen und globalen Modellverständnis führen und ob es Unterschiede bzgl. verschiedener subjektiver Maße im Vergleich zu nicht-personalisierten Erklärungen gibt.

Aufgaben:

- Literaturrecherche (v.a. experimenteller Studien, Journal Paper) im Bereich Personalisierte/Adaptierbare XAI, Counterfactual Explanations, ggf. spezielle Bereiche in der Mensch-KI-Interaktion
- Integration einer eigenen entwickelten Forschungsfrage in den bestehenden Versuchsplan
- Versuchsleitung: Durchführung der Datenerhebung in einer Laborstudie mit Versuchspersonen
- Ausarbeitung des Auswertungsplans, Aufbereitung der Daten und Datenauswertung mittels z.B. SPSS
- Beantwortung der Forschungsfrage und Ableitung relevanter wissenschaftlicher Erkenntnisse

Haben Sie Interesse an dieser Arbeit?
Dann nehmen Sie gerne Kontakt auf:
M. Sc. Lena Kölmel
Tel: 0721 – 608 -44717
Lena.koelmel@kit.edu