Erklärbare und sichere KI

Sommer

(engl. Explainable and safe AI)

Modulnummer: FIN-INF-110496

Kürzel	CP	Semester	Fachsem.	Dauer	Sprache	Niveau
XAI	5	Sommer	ab 3.	1 Semester	deutsch	Bachelor

Link zum LSF:	LSF
Zugang beschränkt:	
Verantwortung:	Sebastian Stober
Dozent:in:	Sebastian Stober
Lehrveranstaltungen:	• Vorlesung *Erklärbare und sichere KI* • Übung *Erklärbare und sichere KI*
Verwendbarkeit:	- B.Sc. INF: Informatik - Wahlpflicht - B.Sc. INF: Studienprofil: Künstliche Intelligenz - B.Sc. CV: Informatik - Wahlpflicht - B.Sc. INGINF: Informatik - Wahlpflicht - B.Sc. WIF: Gestalten und Anwenden - Wahlpflicht - B.Sc. INF (bilingual): Informatik - Wahlpflicht

Angestrebte Lernergebnisse:
Die Studierenden ...

- können Methoden zur Erklärung von Entscheidungen und internen Repräsentationen, sowie Prozessen von Modellen des maschinellen Lernens sicher anwenden
- haben die Fähigkeit, die Qualität von Modellen abseits von Genauigkeitsmetriken einschätzen zu können
- können Modelle zielgerichtet verbessern

Inhalt:
Einordnung unterschiedlicher Kategorien von Erklärbarkeits-Methoden für Modelle des

maschinellen Lernens sowie Anforderungen an Nutzer:innenzentrierte Erklärungen. Darstellung direkt interpretierbarer Modelle (ante-hoc) sowie Analyse- und Visualisierungs-Methoden für Blackbox-Modelle (post-hoc). Zudem werden grundlegende Post-hoc Erklärungstechniken mit modell-agnostischen und modell-spezifischen Ansätzen vorgestellt.

Arbeitsaufwand:

- 56h Präsenzzeit (Vorlesung + Übung)
- 94h selbstständige Arbeit (Vor- und Nachbearbeitung von Vorlesung (OER) und Übung, Bearbeiten von Übungs- und Programmieraufgaben)

Prüfungsvorleistungen:	Studien-/Prüfungsleistungen:	Lehrform / SWS:
	Klausur 120 Minuten Ankündigung der notwendigen Vorleistungen in der ersten Veranstaltungswoche.	• Vorlesung (2 SWS) • Übung (2 SWS)

Voraussetzungen nach Prüfungsordnung:	Empfohlene Voraussetzungen:
keine	Deep Learning für Ingenieure

Medienformen:	Literatur:

Hinweise:

Revision #1
Created 2026-11-20 10:24:26 UTC by Fakultät für Informatik
Updated 2026-11-20 10:24:26 UTC by Fakultät für Informatik