ERKLÄRBARKEIT UND TRANSPARENZ VON KI-METHODEN

10. JUNI 2020 | TAGUNGSBERICHT ZUR WEBKONFERENZ

» Am 10.06.2020 fand im Rahmen der Begleitforschung zum „Innovationswettbewerb Künstliche Intelligenz" des Bundesministeriums für Wirtschaft und Energie die Webkonferenz „Erklärbarkeit und Transparenz von KI-Methoden" statt.

» Vielen KI-Methoden liegen Black-Box-Modelle zugrunde, sodass bei deren Verwendung zumeist nur eingeschränkt nachvollziehbar ist, wie und auf welcher Grundlage ein konkreter Algorithmus eine Entscheidung trifft. Zum Hemmnis für den Einsatz von KI-Methoden kann diese fehlende Nachvollziehbarkeit vor allem in Branchen werden, in denen kritische Entscheidungen getroffen werden. Mehrere Projekte, die im Rahmen des KI-Innovationswettbewerbs gefördert werden, widmen sich daher der Frage, wie eine ausreichende Erklärbarkeit beim Einsatz von KI-Methoden in den jeweiligen Anwendungsgebieten gewährleistet werden kann. Vier der fünf Sprecher:innen stellten im Rahmen der Konferenz konkrete Anwendungen und Umsetzungsmöglichkeiten aus ihren Projekten vor. Im abschließenden Vortrag und der Paneldiskussion wurde dann der Fokus auf eine mögliche Zertifizierung von KI-Systemen gelegt.

» Im ersten Vortrag gab Prof. Marco Huber (Fraunhofer IPA) vom Projekt FabOS eine Einführung in das Thema Erklärbarkeit von KI-Methoden. Er unterschied hier grundsätzlich zwischen zwei Erklärbarkeitstypen und nannte Umsetzungskonzepte und Methoden:

» Globale Erklärbarkeit / Modellerklärbarkeit: Modell "als Ganzes" verstehen

Vorgestellte Umsetzungskonzepte:

- Ante-hoc-Erklärbarkeit: Generieren von erklärbaren White-Box-Modellen auf Basis der Ursprungsdaten (z. B. lineare Modelle, Entscheidungsbäume, Regelsysteme)
- Surrogat-basierte Erklärbarkeit: Generieren von White-Box-Stellvertretermodellen (Surrogat-Modelle) aus Black-Box-Modellen
- Post-hoc-Erklärbarkeit

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

» Lokale Erklärbarkeit/Datenerklärbarkeit: Verstehen einzelner Prognosen auf Datenebene
Genannte Methoden:
- LIME (Local Interpretable Model-Agnostic Explanations)
- Saliency Maps

» Prof. Alexander Löser (Beuth Hochschule für Technik Berlin) vom Projekt Service-Meister stellte Möglichkeiten der Herstellung von Erklärbarkeit für textbasierte Anwendungen vor:
» Lernen aus bekanntem Wissen, automatisierte Suche in langen Textdokumenten, wie z. B. Service-Tickets in der Produktion, die umfangreiches Erfahrungswissen enthalten
» Umsetzungskonzepte:
- Suchmaschinen für medizinische Fragestellungen https://cord19.cdv.demo.datexis.com/: Nicht nur einfache Rückgabe des Artikels, auch automatisierte Ausgabe des Absatzes, je nach Ausgangsfrage, z. B. Behandlungsmethoden (Paragraph Retrieval / Classification)
- Beantwortung von Fragen aus Kontextwissen: https://visbert.demo.datexis.com/

» In seinem Vortrag erläuterte Prof. Philipp Rostalski (Universität zu Lübeck) vom Projekt KI-Sigs, wie mithilfe von Gauß-Prozessen im Medizinbereich nachvollziehbare Algorithmen entwickelt werden können:
» Gauß-Prozesse als Werkzeug für flexible und transparente Machine-Learning-Algorithmen
» Einbeziehung von Vorwissen über parametrisierbare (squared exponential, periodische, nicht-periodische) Kernel-Funktionen
» Vorteile bei Gaußprozess-Regression: Nutzbarkeit von Vorwissen und physikalischen Modellen, Quantifizierung von Unsicherheiten, Nutzung effizienter Algorithmen im Rahmen von Maximum-Likelihood-Schätzung oder Bayes'sche Optimierung

» Im vierten technischen Vortrag wurden von Prof. Urbas (Technische Universität Dresden) vom Projekt KEEN besonderen Anforderungen der Prozessindustrie an Erklärbarkeit sowie zwei Ansätze zur Erzeugung von Transparenz vorgestellt:
» Geringe Datenverfügbarkeit: Teure Messsensorik, Echtzeit-Prozessoptimierung, "lokale" Datenverarbeitung
» Regulierungsanforderungen: Regulator als Zielgruppe für Erklärbarkeit
» Vorgestellte Ansätze:
- Hybride Modellierung: Kombinierte Nutzung von datengetriebenen Black-Box-Modellen, Surrogat-Modellen und mechanistischen White-Box-Modellen
- Kombination datengetriebener und menschlicher Analyseansätze

Gefördert durch:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

» Im fünften Vortrag von Lajla Fetic (Bertelsmann-Stiftung) wurde erläutert, wie KI-Ethik- und Zertifizierungs-Konzepte praktisch umgesetzt werden können:
  » Algo.Rules: Regeln für die Gestaltung algorithmischer Systeme https://algorules.org/
  » Praxisleitfaden zu den Algo.Rules https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/praxisleitfaden-zu-den-algorules-all-1
  » Publikation “From Principles to Practice”: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf
    - enthält Label (in Anlehnung an Energieeffizienzlabels)
    - Anwendungskontext klassifizieren
» In der Paneldiskussion wurden unterschiedliche Herausforderungen und Umsetzungspfade bei der Zertifizierung von KI-Systemen diskutiert. Die KI-Zertifizierung stellt perspektivisch ein Instrument dar, mit dem die Akzeptanz und somit auch der wirtschaftliche Nutzen von KI-Lösungen erhöht werden könnte. Neben der Erklärbarkeit spielen dabei, je nach Anwendung, auch Aspekte wie Zuverlässigkeit und Fairness wichtige Rollen. Im Panel war man sich einig, dass sich wesentliche Herausforderungen für die Festlegung von Schwellenwerten ergeben, die anwendungsbezogen und risikoabhängig festzulegen sein werden. Dennoch wurde die Entwicklung von angemessenen Zertifizierungsmechanismen von allen Teilnehmenden als eine wichtige Aufgabe angesehen, die in den kommenden Jahren vorangetrieben werden muss.

Weitere Informationen zu den KI-Projekten und dem Programm unter
www.ki-innovationen.de

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages