CONSTRUCTING EXPLAINABILITY

TRR 318

Erklärungen gemeinsam entwickeln | 01.2023

Im Newsletter des Transregios (TRR) 318 „Constructing Explainability" präsentieren wir Forschungsprojekte, Workshops zu künstlicher Intelligenz (KI) sowie aktuelle Nachrichten und Vorträge. Sie sind herzlich eingeladen, mit uns auf Twitter zu interagieren und uns eine E-Mail mit Ihrer Frage zu KI zu schreiben. Lassen Sie uns gemeinsam Erklärungen entwickeln!

english version below: click here ▼

Hinter den Kulissen des TRR 318

„Forschung und Management ergänzen sich zwar, aber gehen nicht immer Hand in Hand, haben nicht immer dieselben Ziele und sprechen nicht immer dieselbe Sprache", sagt Ronja Hannebohm. Sie leitet als Geschäftsführerin den Bereich Management im TRR 318. Wie zentral diese Arbeit für die Forschung ist, zeigen ihre vielfältigen Aufgaben. Im Interview berichtet sie von ihrem Berufsalltag und was es heißt, einen großen Forschungsverbund zu organisieren.
Weiterlesen

News

Erster TRR-Schreib-Retreat

Im ersten Writing Retreat sind die Nachwuchswissenschaftler*innen des TRR 318 in Warburg zusammengekommen, um gemeinsam Texte zu schreiben und neue Schreibmethoden zu lernen.

Weiterlesen

Vorträge zu KI und Robotik im Gesundheitswesen

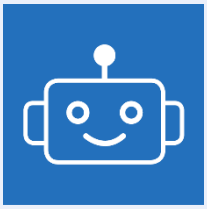
Künstliche Intelligenz kann Mediziner*innen in ihrer Arbeit unterstützen, zum Beispiel bei der Erkennung von Krankheitsbildern auf Röntgenbildern. Im Vortrag am 26.04. zeigt Prof. Philipp Cimiano, welche Bedeutung Erklärungen hierbei zukommt.

Weiterlesen

Digitales TRR-Lexikon

Das neue TRR-318-Wiki funktioniert wie die Wikpedia und enthält alle wichtigen Begriffe aus der Forschung des TRR 318 rund um (Erklärbare) Künstliche Intelligenz sowie Erklärungen und Erklärprozessen. Das Wiki wurde vom Projekt INF entwickelt und wird fortlaufend aktualisiert.

Weiterlesen



Wie Chat-GPT unseren Umgang mit KI verändern kann

Wenn es um KI geht, kommt man derzeit nicht um das Thema ChatGPT herum. Zwei Forschende des TRR mit unterschiedlichen wissenschaftlichen Hintergründen haben sich mit Chatbots beschäftigt und beantworten im Interview Fragen zum Thema:

Soziologe Nils Klowait

Computerlinguist Prof. Dr. Henning Wachsmuth

Weitere News

„ChatGPT und andere generative Modelle suggerieren, dass KI-Systeme einfach ‚funktionieren'. In allen Anwendungsbereichen, in denen eine KI innerhalb eines gegebenen Rahmens zuverlässig und robust eingesetzt werden muss, ist aber Training, Überwachung und Pflege der KI notwendig. Dafür müssen wir verstehen, was Nutzer*innen benötigen und wie sie mit der KI interagieren werden. Wir müssen eine Strategie entwickeln, um effizient Trainingsdaten für die KI zu generieren, bei denen der Mensch die erwartete Lösung vorgibt. Es müssen die richtigen Modelle, Algorithmen und Bewertungsfunktionen für das Training der KI ausgewählt werden. Das trainierte Modell muss auf sogenannte ‚ungesehene' Daten evaluiert werden, um die Robustheit, Generalisierbarkeit und Fehlerhaftigkeit der KI festzustellen, bevor sie in die Anwendung gebracht wird. Im sogenannten Lebenszyklus der KI müssen wir zudem die Performanz der KI konstant überwachen. Die KI muss regelmäßig neu oder weiter trainiert werden, um gewisse Fehler zu vermeiden oder um sie an zeitliche Veränderungen der Daten, dem Anwendungskontext oder der Nutzung anzupassen. Das ist dann ‚Training on the Job' für die KI.“

Prof. Dr. Philipp Cimiano
Stellv. Sprecher des TRR 318 und Informatiker
*Teilprojekte **B01**, **C05** und **INF***

Medientipps

Interview „Erklären ist immer eine zweiseitige Sache“
Wenn sich Menschen einander etwas erklären, beobachten sie oft die Signale von Unverständnis bei ihrem Gegenüber, um ihre Erklärung gegebenenfalls anzupassen. Dieses Verfahren wollen die Wissenschaftler*innen des TRR-Teilprojekts A06 auch in Maschinen ermöglichen. Drei Fragen an die Projektleiterin Hanna Drimalla. **Artikel lesen**

Interview „Algorithmische Diskriminierung: Das Problem von KI und Vorurteilen“
KI-Systeme arbeiten oft mit großen Datenmengen und das schnell. Teilweise sind diese Daten und damit auch die Entscheidungen einer KI personenbezogen, wodurch sich auch Risiken für den Einsatz von KI ergeben. Junior-Professorin Ilona Horwath aus den TRR-Projekten war beim Radio FM4 zu Gast, um über algorithmische Diskriminierung zu sprechen. Der Artikel zum Interview ist online verfügbar. **Artikel lesen**

Video-Reihe zu Künstlicher Intelligenz
Der WDR spricht in ihrer Video-Reihe „Mit KI in die Zukunft“ verschiedene Themen rund um Künstliche Intelligenz an. In den aktuell vier knapp zehnminütigen Videos geht es um die Fragen „Was ist KI?“ und „Was kann KI?“ sowie KI in Medizin und ethischen Fragen zu KI. **Mehr erfahren**

Podcasts zu Künstlicher Intelligenz
In der vierten Folge des Pocasts „Treffen sich Welten“ sprechen eine Informatikerin und ein Schriftsteller über soziale Aspekte der digitalen Transformation. **Mehr erfahren**

Der Podcast „Neuland“ des Hasso-Platter-Instituts behandelt in seiner aktuellen Folge vom 15. Februar erklärbare KI und warum wir diese Form von KI brauchen, damit wir der Technologie besser vertrauen können. **Mehr erfahren**

Veröffentlicht

Forschungsartikel: Verbessern Visualisierungen das Vertrauen in einen Roboter?
Ein Team aus Informatiker*innen der Universität Bielefeld hat in einer Studie untersucht, inwieweit Visualisierungen in Augmented Reality das Verständnis, das Vertrauen und die Akzeptanz von Nutzer*innen gegenüber einem Roboter erhöhen können. Die Ergebnisse finden Eingang in das Forschungsprojekt B05 des Transregios, in dem Erklärungen mittels Visualisierungen gegeben werden sollen. **Weiterlesen**

Forschungsartikel: Die Maschine als Spielerklärerin
Ein und dieselbe Erklärung kann entweder zu ausschweifend oder nicht ausführlich genug sein – je nachdem, wer die Person ist, der etwas erklärt wird. Wie Maschinen dieses Problem meistern und ihre Erklärstrategien individuell und in Echtzeit anpassen können, zeigen Informatiker*innen des TRR 318 in ihrem Modell „SNAPE“. **Weiterlesen**

Forschungsartikel: Wie mentale Vorstellungen die Zusammenstellung von Social-Media-Feeds beeinflussen

Algorithmen organisieren Social-Media-Plattformen und sind zentral für die Anordnung von Beiträgen in den News-Feeds. Dr. Christian Schulz, Medienwissenschaftler im TRR 318, schlägt darauf basierend eine neue Theorie für Social-Media-Plattformen vor, die die Vorstelungen der Nutzer*innen über die Zusammenstellung von Social Media-Feeds miteinbezieht. **Weiterlesen**

Sonderausgabe KI-Magazin: Ein interdisziplinärer Blick auf erklärbare künstliche Intelligenz
Wissenschaftler*innen des TRR 318 präsentieren in einer Sonderausgabe der Zeitschrift für Künstliche Intelligenz des Springer-Verlags ihre neuen Ansätze und Erkenntnisse zu erklärbarer künstlicher Intelligenz. **Weiterlesen**

Gesucht

Teilnehmende für Studien

Eine Auswahl:
- Gesellschaftsspiele erklären und erklärt bekommen,
- sich von virtuellen Agenten anleiten lassen,
- Alltagsbegegnungen mit künstlicher Intelligenz festhalten,
- Robotern Bewegungen beibringen.

Zur Webseite für mehr Informationen zu den aktuellen Studien.
Oder: Mailingliste mit Einladungen zu neuen Studien **abonnieren.**

Schulklassen für KI-Workshops an der Universität Paderborn

Ein TRR-Forschungsteam aus dem Fachbereich Didaktik der Informatik bietet für Schulklassen Workshops zu KI an. Vorwissen ist nicht erforderlich, die Termine werden individuell vergeben. Interessierte Lehrkräfte sind eingeladen, sich per E-Mail anzumelden: **communication@trr318.uni-paderborn.de**

Was habe ich gelernt?

„Bisher habe ich mich für meine Schreibprojekte immer auf den Klassiker Word verlassen. Aber vor allem verrutschende Grafiken und Bilder bereiten in der Anwendung häufig Probleme und stören mich während des Schreibens. Mit der Software zur Textbearbeitung LaTeX hingegen habe ich die Möglichkeit, Formatierung des Textes und den Prozess des Schreibens voneinander zu trennen. In einem LaTeX-Workshop haben wir zum Beispiel mit Overleaf gearbeitet. Hier habe ich zwei Fenster: In einem sehe ich nur meinen Text ohne Bilder, in einem anderen Fenster die Darstellung meines Papers inklusive aller Grafiken, Tabellen und Zitationen. Mit ein paar Befehlen werden Bilder und Grafiken genau an der Stelle einfügt, an der diese sein sollen, und diese Bilder stören mich nicht im Schreibfenster. Ich kann mich ganz auf das Schreiben meines Forschungsartikels oder meiner Dissertation fokussieren.“

Vivien Lohmer
*Doktorandin im **Teilprojekt A04***

TRR digital

Oder direkt per Mail mit Fragen oder Feedback an uns.

Newsletter abonnieren.

Top: german version ▲
Footer: Impressum ▼

Developing explanations together | 01.2023

In the newsletter of Transregio (TRR) 318 "Constructing Explainability" we present our research projects, workshops on artificial intelligence (AI) as well as new publications and upcoming talks. You are invited to interact with us on Twitter and email us with your questions about AI. Let's develop explanations together!

Behind the scenes of TRR 318

"Research and management complement each other, but they don't always go hand in hand, they don't always have the same goals and they don't always speak the same language," says Ronja Hannebohm. As Executive Director, she is in charge of the management area of TRR 318. Her diverse tasks show how central this work is to research. In this interview, she talks about her day-to-day work and what it means to organise a large research network.

Continue Reading

News

First TRR Writing Retreat

In the first Writing Retreat, the doctoral students of TRR 318 met in Warburg to write texts together and to learn new ways of writing.

Read more

Lectures on AI and robotics in healthcare

Artificial intelligence can assist doctors in their work, for example in recognising disease patterns on X-rays. In his lecture on 26 April, Prof. Philipp Cimiano will show the importance of explanations in such cases.

Read more

Digital TRR Dictionary

The new TRR 318 Wiki works like Wikpedia and contains all the important terms from the TRR 318 reseach on (Explainable) Artificial Intelligence as well as explanations and explaining processes. The wiki has been developed by the INF project and will be continuously updated.

Read more

How Chat-GPT can change the way we deal with AI

When it comes to AI, there is currently no way around ChatGPT. In these interviews, two TRR researchers with different scientific backgrounds answer questions about chatbots:

Sociologist Nils Klowait

Computational linguist Prof. Dr. Henning Wachsmuth

More News

“ChatGPT and other generative models suggest that AI systems just 'work.' However, in any application where an AI is to be used reliably and robustly within a given framework, the AI needs to be trained, monitored, and maintained. Therefore, we must understand what users need and how they interact with the AI. We need to develop a strategy to efficiently generate training data for the AI, with humans providing the expected solution. Suitable models, algorithms, and evaluation functions must be selected to train the AI. The trained model must be evaluated on so-called 'unseen' data to determine the AI's robustness, generalisability, and effectiveness before it is deployed. In the so-called life cycle of AI, we also need to monitor the performance of the AI constantly. The AI must be regularly retrained to avoid specific errors or to adapt to temporal changes in the data,

the application context, or the use. This is called 'training on the job' for AI."

Prof. Dr. Philipp Cimiano
Co-speaker of TRR 318 and computer scientist
*Subprojects **B01**, **C05** und **INF***

Published

Research Paper: Do Visualizations Increase Trust in a Robot?
A team of computer scientists from Bielefeld University has conducted a study to investigate how visualizations in augmented reality can increase users' understanding, trust, and acceptance of a robot. The results will feed into the Transregio research project B05, which aims to use visualizations to provide explanations. **Read more**

Research Paper: The machine as a game explainer
The same explanation can be either too detailed or not detailed enough - depending on whom it is described to. In their model "SNAPE", the computer scientists of TRR 318 show how machines can master this problem and adapt their explanation strategies individually and in real-time. **Read more**

Research Paper: How Imaginations Influence the Composition of Social Media Feeds
Algorithms organise social media platforms and are central to the arrangement of posts in news feeds. Based on this, Dr Christian Schulz, media scientist at TRR 318, proposes a new theory for social media platforms that incorporates users' ideas about the composition of social media feeds. **Read more**

AI Journal Special Issue: An Interdisciplinary Take on Explainable Artificial Intelligence
Scientists from TRR 318 present their new approaches and findings on explainable artificial intelligence in a special issue of the Journal of Artificial Intelligence published by Springer-Verlag. **Read more**

What have I learned?

“I have always relied on the classic Word for my writing projects. However, the graphics and images that move around the programme are often a problem and a distraction during the writing process. With the text processing software LaTeX, on the other hand, allows me to separate text formatting from the writing process. For example, in a LaTeX workshop we worked with Overleaf. Here I have two windows: in one window I can see only my text, without any images, and in another window I can see the presentation of my thesis, including all the graphics, tables and quotations. With a few commands, images and graphics are inserted exactly where they should be, and these images do not disturb me in the writing window. I can concentrate on writing my paper or dissertation.”

Vivien Lohmer
*PhD student in **subproject A04***

TRR digital

Or message us directly for questions or feedback.

Subscribe to the newsletter for free.

UNIVERSITÄT PADERBORN

UNIVERSITÄT BIELEFELD

Gefördert durch DFG Deutsche Forschungsgemeinschaft

TRR 318 „Constructing Explainability“

Teilprojekt Ö „Fragen zu erklärbaren Technologien“
Universität Bielefeld
Universitätsstraße 25
33615 Bielefeld

communication@trr318.uni-paderborn.de