KI - Künstliche Intelligenz (2022) 36:207–210
https://doi.org/10.1007/s13218-022-00788-0

EDITORIAL

Explainable AI

Ute Schmid[1] · **Britta Wrede**[2]

Published online: 2 December 2022

During the last years, explainable AI (XAI) has been established as a new area of research focussing on approaches which allow humans to comprehend and possibly control machine learned (ML) models and other AI-systems whose complexity makes the process which leads to a specific decision intransparent. In the beginning, most approaches were concerned with post-hoc explanations for classification decisions of deep learning architectures, especially for image classification. Furthermore, a growing number of empirical studies addressed effects of explanations on trust in and acceptability of AI/ML systems. Recent work has broadened the perspective of XAI, covering topics such as verbal explanations, explanations by prototypes and contrastive explanations, combining explanations and interactive machine learning, multi-step explanations, explanations in the context of machine teaching, relations between interpretable approaches of machine learning and post-hoc explanations, neuro-symbolic approaches and other hybrid approaches combining reasoning and learning for XAI. Addressing criticism regarding missing adaptivity more interactive accounts have been developed to take individual differences into account. Also, the question of evaluation beyond mere batch testing has come into focus.

In this special issue, we address such recent developments in XAI and also take a closer look at interdisciplinary perspectives. We thereby address methods and applications ranging from model change, robotics and image processing over event sequences from a smart home to fact checking. An interdisciplinary insight is given by an analysis of real life explanations of a medical expert in a health care situation.

In our survey "What is missing in XAI so far? An interdisciplinary overview" we bring together interdisciplinary perspectives on explaining and understanding in everyday situations and explaining of AI. We identify challenges and point to open questions regarding faithfulness and consistency of explanations, adaptability to specific information needs and explanatory dialog for informed decision making, the possibility to correct models and explanations by interaction, as well as the need for an integrated interdisciplinary perspective and rigorous approaches to empirical evaluation based on psychological theories.

Johannes Rabold presents in his contribution on a "Neural-symbolic Approach for Explanation Generation based on Sub-concept Detection: An Application of Metric Learning for Low-time-budget Labeling" an approach that enriches visual explanations with verbal local explanations of relational information through combining metric learning and inductive logic programming. It allows a human to provide labels for a small subset of important image parts which are then used to learn a first-order theory that locally explains the black-box with respect to the given image.

Kerzel and colleagues target human–robot interaction scenarios in their contribution on "What's on Your Mind, NICO? XHRI: An eXplainable HRI Framework for Humanoid Robot Social Interaction and Collaboration". Settings that require users to interact with a robot confront users with complex, embodied AI systems that react in unanticipated ways. Explanations about the robot's perception and internal reasoning are, thus, an important prerequisite for smooth and controllable interaction. The authors, therefore, propose an eXplainable HRI (XHRI) approach for efficient communication. With NICO, the Neuro-Inspired COmpanion, an XHRI framework is presented that provides different types of explanations using verbal and non-verbal modalities based on novel approaches to extract such explanations from neural network modules.

In a more abstract scenario Finzel and colleagues in their contribution on "Generating Explanations for Conceptual

✉ Britta Wrede
bwrede3@uni-bielefeld.de

Ute Schmid
ute.schmid@uni-bamberg.de

[1] Cognitive Systems, University of Bamberg, Bamberg, Germany

[2] Systems Engineering for Cognitive Robotics and Cognitive Systems, University of Bremen and Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld, Germany

Validation of Graph Neural Networks: An Investigation of Symbolic Predicates Learned on Relevance-Ranked Sub-Graphs" propose to combine Inductive Logic Programming with Graph Neural Networks (GNN) to generate explanations regarding spatial relations within Kandinsky Patterns with ILP based on the relevance output of GNN explainers and human-expected relevance for concepts learned by GNNs.

Muschalik et al. in their contribution "Agnostic Explanation of Model Change based on Feature Importance" turn our attention to the difficult problem of XAI in the context of online learning in dynamic environments, where models need to be adapted continuously due to the changing data. The authors motivate the problem of explaining model change, i.e. the difference between models before and after adaptation, instead of explaining the models themselves. To this end a first efficient model-agnostic approach to dynamically detecting, quantifying, and explaining significant model changes based on an adaptation of the Permutation Feature Importance measure is provided that allows the explainee to adapt several hyperparameters affecting the explanation.

Hartmann et al. present the project XAINES in their report "XAINES: Explaining AI with narratives". The XAINES project investigates the explanation of AI systems through narratives targeted to the needs of a specific audience, focusing on two important aspects that are crucial for enabling successful explanation: generating and selecting appropriate explanation content, i.e. the information to be contained in the explanation, and delivering this information to the user in an appropriate way. In this article, the project's roadmap towards enabling the explanation of AI with narratives is presented.

With anomalie detection in a smart home the next project report from Baudisch et al. addresses an everyday AI application. Their Project Report "A Framework for Learning Event Sequences and Explaining Detected Anomalies in a Smart Home Environment" presents an approach to explaining detected anomalies in event sequences in a smart home environment based on simple rule analysis. Anomalies are detected with a model learned from the behavior of the residents by finding deviations from the "normal" behavior.

The following report on "Exploring monological and dialogical phases in naturally occurring explanations" by Fisher and colleagues provides an interdisciplinary perspective on naturally occurring explanations. The project investigates the structure of naturally explanations in a medical context and the roles and engagement of the participants with respect to their conversational jobs. By identifying monological and dialogical phases it reveals that despite a high variability in the sequential structure of different explanations there is a pattern indicating that explainees take an active part by initiating dialogical phases. This has consequences for XAI systems as they need to be able to provide similar phases and activities also in interactions with AI systems.

The following article "Towards Explainable Fact Checking" is a summary of a habilitation (doctor scientiarum) thesis submitted by Isabelle Augenstein to the University of Copenhagen. The dissertation addresses several fundamental research gaps within automatic fact checking. The contributions are organised along three verticles: (1) the fact-checking subtask they address; (2) methods which only require small amounts of manually labelled data; (3) methods for explainable fact checking, addressing the problem of opaqueness in the decision-making of black-box fact checking models.

The article "Identification of explainable structures in data with a human-in-the-loop" presents the thesis of Michael C. Thrun which describes the central steps of an approach for a parameter-free methodology for the estimation and visualization of probability density functions which serve as a basis for selecting a hypothesis and an appropriate distance metric independent of the data context. Projection-based Clustering allows for subsequent interactive identification of separable structures in the data. The complete XAI approach is based on a decision tree guided by distance-based structures in data. The DSD-XAI shows initial success in application to multivariate time series and non-sequential high-dimensional data and generates meaningful and relevant explanations that are evaluated by Grice's maxims.

Finally, in the "Interview with the Speakers of the TRR 318 Constructing Explainability Katharina Rohlfing and Philipp Cimiano" Ute Schmid elicits insights into the newly established TransRegio/Collaborative Research Center. In the interview, the two speakers explain the core ideas of their approach to explaining AI and its social meaning in different contexts.

1 Content

1.1 Survey

- *What is missing in XAI sofar? An interdisciplinary overview* [10]
  Ute Schmid and Britta Wrede

1.2 Technical Contributions

- *A Neural-symbolic Approach for Explanation Generation based on Sub-concept Detection: An Application of Metric Learning for Low-time-budget Labeling* [8]
  Johannes Rabold
- *What's on Your Mind, NICO? XHRI: An eXplainable HRI Framework for Humanoid Robot Social Interaction and Collaboration* [6]

Matthias Kerzel, Jakob Ambsdorf, Dennis Becker, Wenhao Lu, Erik Strahl, Josua Spisak, Connor Gäde, Tom Weber and Stefan Wermter

- *Generating Explanations for Conceptual Validation of Graph Neural Networks: An Investigation of Symbolic Predicates Learned on Relevance-Ranked Sub-Graphs* [3]
  Bettina Finzel, Anna Saranti, Alessa Angerschmid, David Tafler, Bastian Pfeifer and Andreas Holzinger

1.3 Project Reports

- *Agnostic Explanation of Model Change based on Feature Importance* [7]
  Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer and Eyke Hüllermeier
- *XAINES: Explaining AI with narratives* [5]
  Mareike Hartmann, Han Du, Nils Feldhus, Ivana Kruijff-Korbayová, Daniel Sonntag
- *A Framework for Learning Event Sequences and Explaining Detected Anomalies in a Smart Home Environment* [2]
  Justin Baudisch, Birte Richter, Thorsten Jungeblut
- *Exploring monological and dialogical phases in naturally occurring explanations* [4]
  Josephine B. Fisher, Vivien Lohmer, Friederike Kern, Winfried Barthlen, Sebastian Gaus, Katharina J. Rohlfing

1.4 Dissertation Abstracts

- *Towards Explainable Fact Checking* [1]
  Isabelle Augenstein
- *Identification of Explainable Structures in Data with a Human-in-the-Loop* [11]
  Michael Christoph Thrun

1.5 Interview

- *Interview with the Speakers of the TRR 318 Constructing Explainability Katharina Rohlfing and Philipp Cimiano* [9]
  Ute Schmid

2 Service

2.1 Conferences and Workshops

The topic of XAI is addressed in numerous workshops—at major AI conferences, such as XXAI@ICML Workshop 2020, or the XAI workshop at IJCAI 2022. The topic of XAI is also covered at the AAAI/ACM Conference on AI, Ethics, and Society (AIES).

2.2 Journals

Several journals presented special issues about XAI in the last years, among them the Artificial Intelligence Journal (Volume 307, June 2022).

2.3 Projects, Organisations, Communities

It is not possible to make a fair selection of ongoing XAI projects, we will only point out two:

- The EU-funded project 'Science and technology for the explanation of AI decision making' https://xai-project.eu
- The DFG Transregional Collaborative Research Centre TRR 318 'Constructing Explainability' https://trr318.uni-paderborn.de

The German Zentrum für vertrauenswürdige KI https://www.zvki.de has a strong focus on comprehensibility and explainability of AI systems for end-user and consumer.

2.4 Resources

An excellent overview over most XAI methods can be found in the book

- Interpretable Machine Learning—A Guide for Making Black Box Models Explainable by Christoph Molnar (last update 2022-11-12): https://christophm.github.io/interpretable-ml-book/.

An in-depth introduction of XAI methods with a focus on visualization methods can be found in the collection

- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds). (2019). Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Augenstein I (2022) Towards explainable fact checking. Künstl Intell
2. Baudisch J, Richter B, Jungeblut T (2022) A framework for learning event sequences and explaining detected anomalies in a smart home environment. Künstl Intell
3. Finzel B, Saranti A, Angerschmid A, Tafler D, Pfeifer B, Holzinger A (2022) Generating explanations for conceptual validation of graph neural networks. Künstl Intell
4. Fisher JB, Lohmer V, Kern F, Barthlen W, Gaus S, Rohlfing KJ (2022) Who does what in the two phases of an explanation?. Künstl Intell
5. Hartmann M, Du H, Feldhus N, Kruijff-Korbayova I, Sonntag D (2022) XAINES: explaining AI with narratives. Künstl Intell
6. Kerzel M, Ambsdorf J, Becker D, Lu W, Strahl E, Spisak J, G"ade C, Weber T, Wermter S (2022) What's on your mind, NICO?. Künstl Intell
7. Muschalik M, Fumagalli F, Hammer B, Hüllermeier Eyke (2022) Agnostic explanation of model change based on feature importance. Künstl Intell
8. Rabold J (2022) A neural-symbolic approach for explanation generation based on sub-concept detection: an application of metric learning for low-time-budget labeling. Künstl Intell
9. Schmid U (2022) Constructing explainability—interdisciplinary framework to actively shape explanations in XAI. Künstl Intell
10. Schmid U, Wrede B (2022). What is missing in AI so far? An interdisciplinary overview. Künstl Intell
11. Thrun C (2022) Identification of explainable structures in data with a human-in-the-loop. Künstl Intell