Interaction with Explanations in the XAINES Project

Mareike Hartmann
Interactive Machine Learning
DFKI
Saarbrücken, Germany
mareike.hartmann@dfki.de

Ivana Kruijff-Korbayová
Multilinguality and Language Technology
DFKI
Saarbrücken, Germany
Ivana.Kruijff-Korbayova@dfki.de

Daniel Sonntag
Interactive Machine Learning
DFKI
Saarbrücken, Germany
daniel.sonntag@dfki.de

***Abstract*—AI systems are increasingly pervasive, and their large-scale adoption makes it necessary to explain their behaviour, for example to their users who are impacted by their decisions, or to their developers who need to ensure their functionality. This requires, on the one hand, to obtain an accurate representation of the chain of events that caused the system to behave in a certain way (e.g., to make a specific decision). On the other hand, this causal chain needs to be communicated to the users depending on their needs and expectations. In this phase of explanation delivery, allowing interaction between user and model has the potential to improve both model quality and user experience. In this abstract, we present our planned and on-going work on the interaction with explanations as part of the XAINES project. The project investigates the explanation of AI systems through narratives targeted to the needs of a specific audience, and our work focuses on the question of how and in which way human-model interaction can enable successful explanation.**

***Index Terms*—XAI, explanations, human machine interaction**

AI systems have huge potential to improve our lives, especially when deployed in high stake scenarios such as healthcare applications or automated driving, where erroneous decisions can have severe consequences [1], [2]. Their impact on human lives comes hand in hand with our need to understand *why* a system behaves in a certain way, to verify that it works as intended, and to estimate the extent to which its decisions can be trusted. In order to enable the use of AI systems in real-world applications, we need to find appropriate ways for explaining their behaviour [3]–[5]. How to do that depends on the audience consuming the model explanations [6]–[8]. For example, *Machine Learning (ML) developers* usually want to test and improve the system, and explanations provide a way of identifying model shortcomings to be fixed [9], [10]. For *domain experts*, such as medical staff or engineers, who use the system for domain-specific applications, explanations serve to improve the co-operation between the domain expert and the machine, e.g. by providing a way of evaluating the reliability of a model's decision.

For both audiences, a central component is the interaction between user and machine based on explanations (see Figure 1), where the model provides an explanation to the user, and the user provides feedback to the model based on the explanation [11]–[13]. For ML developers, providing feedback to the model allows to efficiently fix deficiencies that were identified based on model explanations [10]. For domain experts, the interaction with model explanations benefits the user and the way they use the system: The ability to provide feedback to the model increases user satisfaction [11], [14], [15] and their trust in the system [16]. Finally, the social sciences point out that explanations themselves should be an interactive communication between the model as explainer and the user as explainee [17], [18].

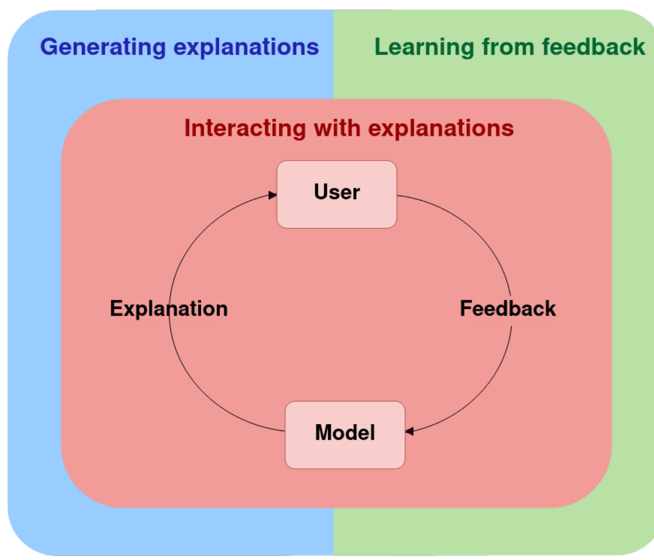

Fig. 1. Interaction with explanations (middle part) plays a central role for XAI, which requires the generation of model explanations (left part) and the integration of user feedback (right part).

The goal of our work presented here is to deliver explanations in an interactive loop that aligns with a target audience's needs. We investigate this task as a part of the XAINES project[1], that aims at explaining AI systems through *narratives*, i.e. an event is explained by giving an account of the events that caused it [19]. Figure 1 shows an overview over the different research areas involved in our task. In the following, we outline our on-going and planned work on explanation generation (Section I) and the interaction with explanations (Section II).

[1]https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/projekt/xaines/

I. Generating Explanations

In addition to targeting two different audiences, XAINES distinguishes two types of explanations (see Figure 2). *ML narratives* convey the causal chain leading to a model prediction, and can primarily be used to improve the model. For example, saliency maps as ML explanations [20] can reveal that a model picks up on irrelevant features to classify X-ray images [21]. *Domain narratives* describe sequences of domain-specific events that led to a specific outcome, and can e.g. be used by domain experts to assess if a model decision is justified. We explore the generation of both types of explanations in the context of describing visual content, with a focus on providing explanations for systems used in the medical domain, e.g. for speech-based image annotation [22] or medical decision support [23].

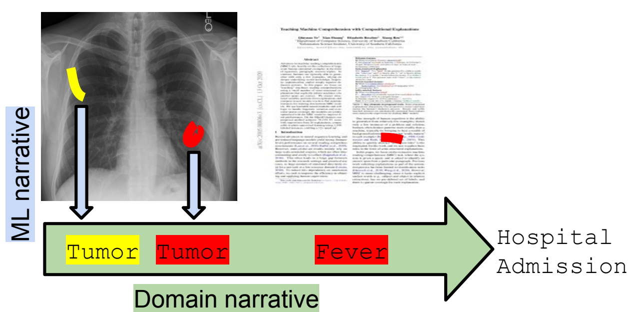

Fig. 2. Examples of ML and domain narratives for a medical decision support system.

A. Information Extraction from Images

The generation of domain narratives requires the extraction and description of relevant information from domain-specific data in various forms, such as X-ray images or health records [24]. For domain narratives, we focus on the tasks of image captioning [25]–[27] and visual story telling [28], i.e. the description of relevant information in an image or sequences thereof, and use saliency methods such as Grad-CAM [29] to give ML explanations for the generated descriptions and classifier decisions, e.g. in the context of skin cancer recognition [30], [31]. The underlying research questions are if image descriptions are suitable as domain narratives, how their interplay with ML explanations impacts the explanation process, and how to best generate relevant narratives for visual or multimodal content. In [32], we propose an image captioning model that conditions generation on selected visual information to model the fact that humans restrict their explanation of an event to a subset of selected causal connections [18].

II. Interacting with Explanations

For explanation delivery, we focus on making use of interaction between user and machine: First, we investigate how visual explanations can be delivered in an explanation-feedback loop, that aims at improving the model based on human feedback, and allows personalization of explanations. Second, we explore how to move beyond a one-way broadcast of explanation content by modelling explanation as a conversational interaction between user and machine.

A. Interaction with Visual Explanations

We want to enable interaction with visual explanations of classifier decisions in the Interactive Machine Learning (IML) framework, where models are improved based on feedback gained from interaction with users. Building on related work exploring the explanation-feedback loop [12], [13], [33], we will address the open questions of the best mechanism for integrating feedback into the model [34], the type of feedback that is most helpful for model improvement, and how to best evaluate the framework, either in terms of model accuracy, or in terms of user-centric metrics. In addition to ML explanations, we ask if IML methods can also be used for rendering domain narratives. We plan to gain first insights based on simulated feedback, and to then consolidate findings in an interactive user study. Along with providing a means for general model improvement, the interaction between user and model can be exploited to adapt explanations, e.g. as personalized image descriptions. Our experiments in [32] show promising initial results for caption personalization using interactive re-ranking of decoder output, which we plan to explore further in the future.

B. Conversational Interaction as Narrative Explanation of AI

Human explanations are interactive and incremental, allowing participants to challenge, query, negotiate, discuss and clarify the explanation content, ideally until mutual understanding and agreement is achieved [35]. We aim at modelling this important aspect of explanation as a goal-oriented dialog between the user and the machine, where the goal is to achieve mutual understanding with respect to the explanation. We envision the dialog system to be adaptive with respect to the user, as the amount of detail of the explanatory dialogue should be conditioned on their abilities and expectations [18]. Oversimplified explanations that lead to unjustified trust must be avoided [36], therefore one challenge is to find a trade-off between persuasive and descriptive explanation strategies [37]. Other challenges include how to best present the narrative, e.g. by splitting it into multiple installments [38], and how to adapt user representations over time. We are planning to investigate these research questions using a multimodal interactive explanation use case in a Motion Synthesis framework, focusing on urban street scenes. The proposed dialog system should also adapt to user intent, by matching a user query with an appropriate explanation method. A query like *Which inputs contributed most to model output?* matches with an explanation method highlighting parts of the input, e.g. based on input gradients [39]. In contrast, a query like *What (general) patterns in the (training) data are responsible for an output?* matches with an explanation resulting from a probing task [40]. For matching intent to explanation, we plan to explore standard intent classification [41], [42] and textual similarity models [43], [44].

ACKNOWLEDGMENTS

The research was funded by the XAINES project (BMBF, 01IW20005) and the pAItient project (BMG, 2520DAT0P2).

REFERENCES

[1] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[2] Thomas Alexander Sick Nielsen and Sonja Haustein. On sceptics and enthusiasts: What are the expectations towards self-driving cars? *Transport policy*, 66:49–55, 2018.

[3] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6):542–558, 1981.

[4] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.

[5] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.

[6] Roberta Calegari, Giovanni Ciatto, Jason Dellaluce, and Andrea Omicini. Interpretable narrative explanation for ml predictors with lp: A case study for xai. In *WOA*, pages 105–112, 2019.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[8] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021.

[9] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5-6):413–451, 2020.

[10] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *arXiv preprint arXiv:2104.15135*, 2021.

[11] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 82–91, 2007.

[12] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[13] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

[14] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[15] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[16] Maartje M. A. de Graaf and Bertram F. Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*, pages 19–26. AAAI Press, 2017.

[17] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[18] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[19] Stephen P Norris, Sandra M Guilbert, Martha L Smith, Shahram Hakimelahi, and Linda M Phillips. A theoretical framework for narrative explanation in science. *Science Education*, 89(4):535–563, 2005.

[20] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[21] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.

[22] Daniel Sonntag, Christian Schulz, Christian Reuschling, and Luis Galarraga. Radspeech's mobile dialogue system for radiologists. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 317–318, 2012.

[23] Alexander Prange, Michael Barz, and Daniel Sonntag. Speech-based medical decision support in vr using a deep neural network. In *IJCAI*, pages 5241–5242, 2017.

[24] Daniel Sonntag and Hans-Jürgen Profitlich. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial intelligence in medicine*, 93:13–28, 2019.

[25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[27] Luke Melas-Kyriazi, Alexander M Rush, and George Han. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 757–761, 2018.

[28] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[30] Duy MH Nguyen, Thu T Nguyen, Huong Vu, Quang Pham, Manh-Duy Nguyen, Binh T Nguyen, and Daniel Sonntag. Tatl: Task agnostic transfer learning for skin attributes detection. *arXiv preprint arXiv:2104.01641*, 2021.

[31] Fabrizio Nunnari, Md Abdul Kadir, and Daniel Sonntag. On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 241–253. Springer, 2021.

[32] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. *KI-Künstliche Intelligenz*, 34(4):571–584, 2020.

[33] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1):1064–1074, 2019.

[34] Active learning in image captioning. https://iml.dfki.de/active-learning-in-image-captioning/.

[35] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1033–1041, 2019.

[36] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[37] Bernease Herman. The promise and peril of human evaluation for model interpretability. *arXiv e-prints*, pages arXiv–1711, 2017.

[38] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[40] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.

[41] Chunxi Liu, Puyang Xu, and Ruhi Sarikaya. Deep contextual language understanding in spoken dialogue systems. In *Sixteenth annual conference of the international speech communication association*, 2015.

[42] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and S Yu Philip. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267, 2019.

[43] Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, 2018.

[44] Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. A hybrid neural network model for commonsense reasoning. *EMNLP 2019*, page 13, 2019.