Explainable AI through Rule-based Interactive Conversation

Christian Werner
Christian.Werner@viadee.de

ABSTRACT

This is a work-in-progress paper which proposes a rule-based, interactive and conversational agent for explainable AI (XAI) called ERIC. It includes research from XAI, human computer interaction and social science to provide selected, personalized and interactive explanations.

KEYWORDS

explainable, artificial intelligence, conversational, agent

1 INTRODUCTION

Nowadays, artificial intelligence (AI) has an ubiquitous impact on our life. This involves product recommendations, risk assessment and systems that are essential for people's survival such as medical diagnosis systems. Especially in case of such critical decisions being made by a system, the question arises why and how it came to a specific decision [3]. The problem is that many of the underlying algorithms of such systems appear as a black-box to the user and therefore suffer in terms of transparency [1]. This is the driver for the research field of so-called explainable AI (XAI). It provides a set of methods which can be used to describe the behaviour of a machine learning (ML) model and as such provides a certain degree of transparency [1]. The current research focuses on the development of new and mostly isolated XAI methods, such as Surrogate Models, Partial Dependency Plots, or Accumulated Local Effects rather than on what really makes up a good overall approach to explain a model's behaviour to the user [10]. The research question is how the results of such methods can be used to answer the questions humans have about ML decision making? This work-in-progress paper introduces a new XAI system called ERIC - a **R**ule-based, **I**nteractive and **C**onversational agent for **E**xplainable AI. ERIC applies the most popular XAI methods on a ML model to extract knowledge that is stored within a rule-based system. A potential user can communicate with ERIC through a chat-like conversational interface and receive appropriate explanations about the ML model's reasoning behaviour. This system is specifically targeted to domain experts and seeks to provide everyday explanations. It combines insights from the research fields of AI, human computer interaction and social science [12]. Other than existing related conversational system (e.g. the Iris agent for performing data science tasks [2], or the LAKSA agent for explaining context-aware applications [8]), ERIC focuses on the explanations of ML models.

2 METHODOLOGY

Research proposed in this paper follows a Design Science Research (DSR) approach that is aimed to iteratively elaborate requirements, implement and test them with real users. Requirements are drawn up from theoretical investigations in literature, existing solution approaches and findings from user experiments.

3 PRELIMINARY RESULTS

System goals. Trust is essential when humans communicate with a system and driver for XAI [12]. However, to generate trust, an XAI system must first and foremost provide transparency regarding its decision making process [13]. Furthermore, the system must present information in an understandable manner and avoid inconsistencies within the information it presents [15].

Intelligibility types. Intelligibility types describe a set of intelligible elements which form a query paradigm that is derived from questions users of intelligent systems often ask [6]. Results from various experiments hint that these question-answer constructs can help to build mental models of a system in a user's mind who can then develop a certain level of trust regarding the system's reasoning [9] [5]. Among others, ERIC implements the following intelligibility types: Why, Why-not, What-if, How-to. Suitable explanations such as rule-based explanations, feature attributions and counterfactual explanations are used as output.

Provide selected explanations. Selecting the right explanations for a context is one of the major challenges for an XAI agent. Not every explanation type is suitable to answer a user issued question and not every XAI method is applicable in every situation [7]. Thus, ERIC includes specific domain knowledge about when to present what type of explanation based on contextual factors.

Provide personalized explanations. Explanations provided to a user must be tailored to the specific need and interest of the user. This involves the complexity of the explanations (number of elements), the prioritization of information (which elements are important for the user), and the presentation format (textual vs. visual) [14]. ERIC seeks to personalize explanations for a user by extracting preferences from user actions and by direct information elicitation.

Provide interactive explanations. One of the main insights about explanations from social science is that an explanation naturally happens in an interactive conversation [11]. Hence, a user should have the possibility to actively explore the underlying ML model as a continuous process. By doing that, the user can develop step-by-step trust in the system [4]. ERIC implements a dialogue model that enables the user to iteratively query different types of information. The presentation of an explanation is never an end point and allows for further inquiries.

4 STATE AND FUTURE DIRECTIONS

A first prototype of ERIC is implemented using the rule-based programming language CLIPS and a Python interface revealing promising results. The prototype allows for a basic interaction about a Python-based ML model using the proposed intelligibility types. Further requirements need to be elaborated and implemented to further specify ERIC's capabilities. User testing is essential to validate the effectiveness of ERIC and is still pending. An online available prototype is being planned.

REFERENCES

[1] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

https://doi.org/10.1109/ACCESS.2018.2870052

[2] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 473.

[3] Leilani Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[4] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining Explanation For "Explainable Ai". In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 197–201.

[5] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*. https://doi.org/10.1109/VLHCC.2013.6645235

[6] Brian Y Lim. 2012. *Improving understanding and trust with intelligibility in context-aware applications*. Ph.D. Dissertation. figshare.

[7] Brian Y Lim and Anind K Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 13–22.

[8] Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. ACM, 157–166.

[9] Brian Y Lim and Anind K Dey. 2013. Evaluating intelligibility usage and usefulness in a context-aware application. In *International Conference on Human-Computer Interaction*. Springer, 92–101.

[10] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. *CoRR* abs/1903.02409 (2019). arXiv:1903.02409 http://arxiv.org/abs/1903.02409

[11] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1033–1041.

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

[13] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[14] Johannes Schneider and Joshua Peter Handali. 2019. Personalized explanation for machine learning: A conceptualization. (2019).

[15] Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots.