Explanation through Dialogue for Reasoning Systems

Doctoral Consortium

Yifan Xu
Department of Computer Science, The University of Manchester
Manchester, United Kingdom
yifan.xu@manchester.ac.uk

ABSTRACT

Explainability and transparency are becoming more critical in logical reasoning, such as in self-driving cars and medical care, where poor decisions can cause harm and, in the worst situations, death. To ensure such systems are morally sound, reliable, and secure, they must be capable of explaining their output or procedures in a human-understandable way. In this research, a dialogue explanation framework for Rule-based reasoning systems is presented to identify and explain discrepancies between the user and the system. It allows the system to explain itself by simply asking and answering "Why?" and "Why not?" questions. The formal properties of this framework and a small user evaluation that contrasts dialogue-based explanations with the proof trees generated by the reasoning system are described.

KEYWORDS

Machine Reasoning; Knowledge Representation; Rule-based Reasoning; Explanation

ACM Reference Format:
Yifan Xu. 2023. Explanation through Dialogue for Reasoning Systems: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Humans are supported by autonomous agents in a variety of services and fields, with differing levels of intelligence and autonomy. It's crucial to make sure these systems are morally sound, reliable, and secure. The use of explainability is a newly developed method to help address these issues. Reasoning is an essential part of the human explainable capacity, which is the process of combining knowledge and beliefs to make new conclusions [7]. Automated expert systems, also known as rule-based reasoning systems which typically work by asking users to respond to a series of questions, have become available to users for giving advice and direction in certain specified subjects [10]. Some of them, such as MYCIN, can explain their reasoning to users [3]. This means it can describe its reasoning steps: how a request for data is related to a goal, how one goal leads to another, and how a goal is achieved. However, such explanations have limitations. When the system contains a large number of rules and facts, the explanation will be complex and the user will find it hard to follow. Explanation through dialogue can be an understandable solution to address this issue.

A dialogue framework has been developed to explain the behavior of a system programmed using the BDI (Beliefs-Desires-Intentions) paradigm which has many similarities to rule-based reasoning systems [5]. It defines a turn-based system and allows users to ask questions about reasons behind the selection of plans of action within the system, but does not provide a way to explain deductive reasoning (which is our focus).

This research will concentrate initially on describing hand-crafted rule-based reasoning systems, and later on AI systems with learned rules. It aims to explore the use of dialogue explanations for such systems and empower the system to perform collaborative explanations. The explanation through dialogue mechanism has two main stages. In the first stage, the dialogue mechanism only delivers an explanation when the user disagrees with the system's conclusions. The next stage is the dialogue theory that provides explanations for both agree and disagree situations. The system framework mainly consists of reasoning deduction and dialogue explanation production. A proof tree (seen Fig. 1) will represent the reasoning deduction.

2 PROPOSED WORK

As a start point, a rule-based reasoning system (called the Covid Advice System) has been implemented using the Prolog language with the first stage of dialogue mechanism, followed by a user evaluation. As mentioned above, the work that has been done so far is mainly about the first stage of the dialogue mechanism, and only has discussed disagreeing situations where the user's information is different from that possessed by the system has been considered. The user evaluation aims to use the user's ability to discover this mismatch following the explanatory process as one of our metrics for assessing the utility of the explanation.

There are two questions that this research is trying to answer. Can dialogue explanation provide an understandable explanation for Rule-based Reasoning systems? Can dialogue explanation provide an understandable explanation for an AI system with learned rules? The work that has been done so far only could answer the first research question.

2.1 Framework

The Covid Advice system is a Rule-based Reasoning system with a set of rules $\boldsymbol{R}$ which are deployed to reason with knowledge-based facts $\boldsymbol{F}$. Its dialogue mechanism focuses on the disagreeing situation and performs a 'one step' dialogue explanation for any particular "why do you believe it" or "why not do you believe it?" question.

The starting point is there are two players: the system and the user. We assume the system has drawn a conclusion generated using reasoning deduction and represented it as a proof tree. If

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

the user disagrees with the conclusion, the task of the dialogue is to identify the cause of the difference between these two players. Fig. 1 shows a simple deduction of the conclusion Jack can meet his friends.

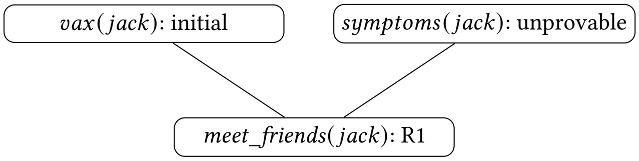
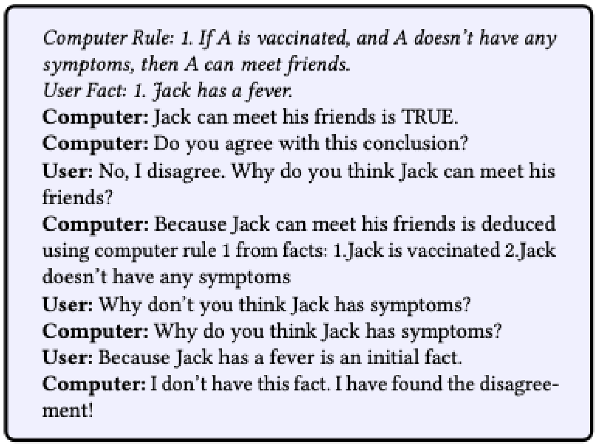


Figure 1: A Proof Tree showing why Jack can meet his friends using $R1 : \{vax(X), \neg symptoms(X)\} \rightarrow meet\_friends(X)$. R1: You can meet friends if you have been vaccinated and display no symptoms, and the initial fact set $\{vax(jack)\}$ means Jack is vaccinated.

2.2 Dialogue

This refers to the example in Figure 1: If a user disagrees, for instance, they know Jack has a fever. Once the disagreement occurs, the user could ask why a particular node is believed or why a conclusion is unprovable. For each why question, the system would provide a 'one-step' explanation giving the last rule used to make the deduction for avoiding redundancy (Fig. 2), and prompts the user to ask a set of follow-up 'Why?' or 'Why not?' questions regarding each piece of information. While for each 'why not' question, the system could flip this around and ask the user why they believe it. The user's answer can then help the system identify where the confusion or disagreement lies and achieve a better understanding.



Computer Rule: 1. If A is vaccinated, and A doesn't have any symptoms, then A can meet friends.
User Fact: 1. Jack has a fever.
Computer: Jack can meet his friends is TRUE.
Computer: Do you agree with this conclusion?
User: No, I disagree. Why do you think Jack can meet his friends?
Computer: Because Jack can meet his friends is deduced using computer rule 1 from facts: 1.Jack is vaccinated 2.Jack doesn't have any symptoms
User: Why don't you think Jack has symptoms?
Computer: Why do you think Jack has symptoms?
User: Because Jack has a fever is an initial fact.
Computer: I don't have this fact. I have found the disagreement!

Figure 2: Dialogue Explanation Example

2.3 User Evaluation

A user evaluation was conducted to reveal the performance of the dialogue mechanism with the Covid Advice System. It comprised 24 volunteers who were staff and students from the Department of Computer Science at the University of Manchester. We hypothesize that when a user doesn't understand or disagrees with the computer's conclusion, the dialogue explanation with reasoning will help them identify if either they have different rules, or if there are facts that they do not have. In the SWI-Prolog terminal interface, participants were presented with two scenarios out of a possible six. Each scenario was completed by the same number of participants and followed by a short questionnaire. Out of 24 responses, 83.3% preferred dialogue explanation to the proof tee explanation, 18 (75%) said it was easy to understand the explanation.

3 RELATED WORK

Early rule-based expert system [16] explanations focused in particular on explanation framework [3, 20], concept [8, 13, 18], and the human-computer interface (HCI) through which the explanation was supplied [11, 17]. The most sophisticated ways for explanation involve an "intelligent" conversation with the system user that is done in simple terms and using interactive methods [6]. However, little progress has been achieved in these early explanations. Few of these could ensure users really understood the content of the explanation. In particular, as the rule-chaining process became more complex, their explanations became increasingly difficult to follow [9].

Argumentation is becoming one of the main reasoning methods to enable explainability in many AI techniques [19], including classifiers [4, 14], knowledge-based systems [1], and AI planning [12]. Bex and Walton [2] utilize argumentation models in dialogue and enable the explainee to question and dispute the provided explanations [11]. Despite the natural affinity of argumentation models to dialogues [2], few such dialogue explanation models have been evaluated in the real human case study.

4 FUTURE WORK

In the future, this research will focus on answering the second research question: Can dialogue explanation provide an understandable explanation for an AI system with learned rules? To achieve this, an AI system will be implemented by expanding our original Rule-based Reasoning system with learned rules. These learned rules will be extracted from machine learning models using the REM algorithm. REM extracts rules from Deep Neural Networks (DNN), which offers a potential opportunity to explore the explanation for such a system [15]. Meanwhile, the second stage of the dialogue mechanism will be modified, which enables the use of dialogue to provide explanations for both agree and disagree situations. To measure whether dialogue explanation can provide an explanation to understand, an exploratory user evaluation will be carried out.

There is a large body of literature on interpretable machine learning, focused on visualization and providing a single explanation to all users. A dialogue system assumes that an explanation is a collaborative process in which the system determines what information it is that the user wants. Dialogue explanation with reasoning allows the user and system to co-create an explanation based on the user's content. This viable mechanism empowers machines with the human ability to explain their actions. It also offers a significant opportunity to further our knowledge of the conversation approach to explainable AI.

REFERENCES

[1] Abdallah Arioua, Nouredine Tamani, and Madalina Croitoru. 2015. Query answering explanation in inconsistent datalog+/- knowledge bases. In *Database and Expert Systems Applications*. Springer, 203–219.

[2] Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* 7, 1 (2016), 55–68.

[3] William J Clancey. 1983. The epistemology of a rule-based expert system—a framework for explanation. *Artificial intelligence* 20, 3 (1983), 215–251.

[4] Oana Cocarascu, Andria Stylianou, Kristijonas Čyras, and Francesca Toni. 2020. Data-empowered argumentation for dialectically explainable predictions. In *ECAI 2020*. IOS Press, 2449–2456.

[5] Louise A Dennis and Nir Oren. 2021. Explaining BDI agent behaviour through dialogue. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[6] Armin Fiedler. 2001. Dialog-driven adaptation of explanations of proofs. In *International Joint Conference on Artificial Intelligence*, Vol. 17. Citeseer, 1295–1300.

[7] Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science* 4, 1 (1980), 71–115.

[8] Carmen Lacave and Francisco J Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17, 2 (2002), 107–127.

[9] Carmen Lacave and Francisco J Diez. 2004. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review* 19, 2 (2004), 133–146.

[10] Stuart Leader, Reinhold Behringer, Ah-Lian Kor, and Nick Cope. 2016. A Rule-Based Guidance (RBG) system with graphical representation of uncertainty. In *2016 Future Technologies Conference (FTC)*. IEEE, 632–636.

[11] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409* (2019).

[12] Nir Oren, Kees van Deemter, and Wamberto W Vasconcelos. 2020. Argument-based plan explanation. In *Knowledge Engineering Tools and Techniques for AI Planning*. Springer, 173–188.

[13] James A Reggia and Barry T Perricone. 1985. Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine* 15, 4 (1985), 161–167.

[14] Naziha Sendi, Nadia Abchiche-Mimouni, and Farida Zehraoui. 2019. A new transparent ensemble method based on deep learning. *Procedia Computer Science* 159 (2019), 271–280.

[15] Zohreh Shams, Botty Dimanov, Sumaiyah Kola, Nikola Simidjievski, Helena Andres Terre, Paul Scherer, Urska Matjasec, Jean Abraham, Mateja Jamnik, and Pietro Liò. 2021. REM: An Integrative Rule Extraction Methodology for Explainable Data Analysis in Healthcare. *medRxiv* (2021).

[16] Edward H Shortliffe, Stanton G Axline, Bruce G Buchanan, Thomas C Merigan, and Stanley N Cohen. 1973. An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Computers and Biomedical Research* 6, 6 (1973), 544–560.

[17] Rudi Studer, V Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: principles and methods. *Data & knowledge engineering* 25, 1-2 (1998), 161–197.

[18] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.

[19] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021).

[20] Michael R Wick and William B Thompson. 1992. Reconstructive expert system explanation. *Artificial Intelligence* 54, 1-2 (1992), 33–70.