Dialogue-Based Explanations of Reasoning in Rule-based Systems

Yifan Xu[1][0000−0003−2303−1531], Joe Collenette[2][0000−0001−6179−2038], Louise A. Dennis[3][0000−0003−1426−1896], and Clare Dixon[4][0000−0002−4610−9533]

Department of Computer Science, The University of Manchester, United Kingdom
{yifan.xu,joe.collenette,louise.dennis,clare.dixon}@manchester.ac.uk

Abstract. The recent focus on *explainable artificial intelligence* has been driven by a perception that complex statistical models are opaque to users. Rule-based systems, in contrast, have often been presented as self-explanatory. All the system needs to do is provide a log of its reasoning process and its operations are clear. We believe that such logs are often difficult for users to understand in part because of their size and complexity. We propose dialogue as an explanatory mechanism for rule-based AI systems to allow users and systems to co-create an explanation that focuses on the user's particular interests or concerns. Our hypothesis is that when a system makes a deduction that was, in some way, unexpected by the user then locating the source of the disagreement or misunderstanding is best achieved through a collaborative dialogue process that allows the participants to gradually isolate the cause. We have implemented a system with this mechanism and performed a user evaluation that shows that in many cases a dialogue is preferred to a reasoning log presented as a tree. These results provide further support for the hypothesis that dialogue explanation could provide a good explanation for a rule-based AI system.

Keywords: Explainable Artificial Intelligence · Machine Reasoning · Rule Based · Artificial Intelligence · Dialogue For Explanation.

1 Introduction

Explainable Artificial Intelligence is increasingly gaining attention in safety-critical domains such as self-driving cars and medicine, where a possible wrong decision might result in people dying [6]. The concern is that most AI systems are unable to explain the reasoning behind their actions or decisions, which can cause significant issues, particularly when agents propose incorrect actions or give inappropriate advice. The need to equip such systems with human-like explainable abilities has grown [9].

Reasoning, which is the process of synthesising facts and beliefs to make new decisions, is a fundamental component of humans' explanatory mechanisms [7]. In the 1970s, rule-based expert systems, used knowledge-based reasoning to make recommendations based on user-provided answers [15]. Some of these had the

capability to provide explanations. MYCIN [16], for instance, could extract an explanation from its current decision to explain why it was requesting additional information from the user. Other systems tried to offer simple but comprehensible reports about the domain and the reasoning [17, 11]. Few of these could ensure users really understood the content of the explanation. When the rule-chaining process of such system becomes very complex, their explanations are difficult to follow [8].

Argumentation can be used to solve a variety of real-world issues, including producing explanations. It can provide explanations that are more closely aligned with human thought processes [19]. Argumentation-based explanation could be used in explaining various models, such as recommendation systems [3, 12], classification [14, 4], probabilistic methods [18], decision making [2], knowledge-based systems [1], planning [10] and logic systems [13]. Dialogue-based explanations may also be seen as an issue that the argumentative strategy addressed on a practical level [19]. Dennis and Oren's [5] dialogue-based technique is intended to explain the behaviour of a system programmed using the *Beliefs-Desires-Intention* logic-based programming paradigm which has many similarities to classical Rule-based systems. This technique defines a turn-based system which enables a user to ask questions about the reasons behind the selection of plans of action within the system. The claim is that this assists the user in comprehending why or, more importantly, why not a certain action was taken. However this claim is not evaluated within the paper, nor does their approach address the use of deduction in reasoning about beliefs and goals. This paper takes the work of Dennis and Oren as a starting point and aims to apply it to the use of dialogue to provide explanations for logical deductions in rule-based systems.

Our framework consists of two steps: rule-based deduction and dialogue statement generation for explaining the deduction. In some situations, particularly when explaining *why not*, these steps need to be interleaved. We represent deductions as trees where the nodes represent facts that have been deduced. The dialogue system allows the user to ask 'why' or 'why not' questions about nodes within the tree and provides a 'one step' explanation for any question in terms of the last rule used to deduce the fact or alternatively (in the case of 'why not' questions) can interrogate the other dialogue participant about why they believe something to be the case. Follow up questions can then be asked to narrow down any source of disagreement or confusion.

Our research question is: Can dialogue provide an understandable explanation for rule-based reasoning? We measure understandability by how easy it is for a user to locate the cause of a disagreement between themselves and the system.

2 Framework

We have a language of terms, $\mathcal{L}$, defined in a standard way, and a set of labels L which include two special labels, *initial* and *unprovable*. Our rule-based system consists of a set of initial facts, F, of positive literals in $\mathcal{L}$; and a set of rules,

R. A rule is a Horn clause consisting of a non-empty set of literals in $\mathcal{L}$ (the antecedents, A), and a consequent, a positive literal $C \in \mathcal{L}$, and a label $l \in L \backslash \{initial, unprovable\}$. We write a rule as $l : A \rightarrow C$. We assume that labels in R are unique – i.e., there is only one rule labelled l in any set of rules, R. Backward-chaining deduction with negation as failure is performed in the standard Prolog way to check whether some literal, l, follows from F and R. We have implemented this system in Prolog and the results of deduction are stored as trees, where a node in the tree is a pair of a positive literal, l and a label. Literals in F are labeled *initial*, and literals that can not be deduced from F using the rules in R are labelled *unprovable*. All other nodes are labelled with the label of the rule used to deduce that node. A node's parents in the tree are the facts that made the antecedents to the rule used to deduce that node true.

We have developed a Covid Advice System with a number of example sets of rules and facts based around Covid-19 restrictions which was used in our evaluation[1]. To show how the system works, Fig. 1 shows a simple deduction of the conclusion that Jack can meet friends using $Rule1 : \{vaccinated(X), \neg symptoms(X)\} \rightarrow can\_meet\_friends(X)$, and the initial fact set $\{vaccinated(jack)\}$

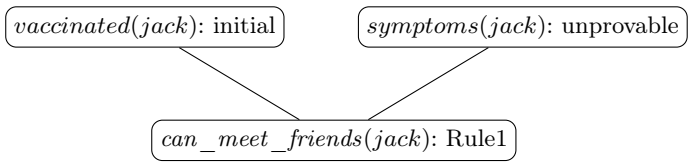


Fig. 1: A Proof Tree

We focus our attention on situations where a user has a slightly different set of rules or facts to the system and so disagrees with the conclusion drawn by the system. The task of an explanation, in this situation, is to identify the cause of the disagreement between user and system.

2.1 Dialogue

We extend our Covid Advice System with a turn-based dialogue system. Once a disagreement occurs, a user can ask why a particular node is believed or why a conclusion is unprovable. For each specific why question, the system provides a 'one-step' explanation giving the last rule used to make the deduction and the facts that made the antecedents of that rule true. It then prompts the user with a set of possible follow-up questions in the form 'Why?' or 'Why Not?' depending upon whether the antecedent is a positive or negative literal. If the user asks a why not question, the system switches to asking questions itself and asks why the user believes this fact to be true. As this process continues it is sometimes

[1] Our system can be found at https://github.com/xyfLily/Rule-based-system

necessary for the computer to perform new deductions in order to explain why rules with negative literals as antecedents did not apply[2].

A sample dialogue in our system is shown in Fig. 2 where the user is using a rule that if Person A has a negative result from a PCR test, then this means Person A has taken a Covid-19 test, while the computer does not have this rule. The computer is therefore advising the user to self-isolate on the grounds they have been in contact with someone who has Covid-19 and, although they are vaccinated, they have not yet taken a Covid-19 test themselves.

Computer Rule: 1. If A has been pinged, and A is vaccinated, and A has not taken a test, then A needs to self-isolate.
User Rule: 4. If A is negative in the PCR test then A has taken a test
Computer: Sarah needs to self isolate is TRUE.
Computer: Do you agree with this conclusion?
User: No, I disagree. Why do you think Sarah needs to self isolate?
Computer: Sarah needs to self isolate is deduced using computer rule 1 from facts: 1.Sarah has been pinged 2.Sarah has been vaccinated 3.Sarah has not taken a test
User: Why do not you believe Sarah has taken a test?
Computer: Why do you believe Sarah has taken a test?
User: Because of user's rule 4.
Computer: I don't have this rule 4. I have found the disagreement!

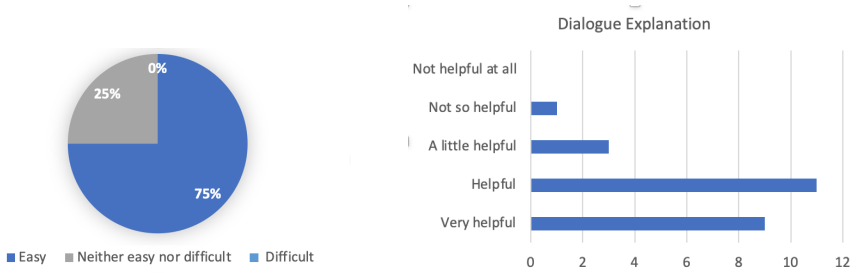
Fig. 2: Example of a dialogue explanation where the user and computer disagree on whether Sarah needs to self isolate.

3 User evaluation

A user evaluation was conducted to reveal the performance of the dialogue mechanism as an explanatory mechanism in comparison with the deduction graphs produced by 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 hypothesized that when a user did not understand or disagreed with the computer's conclusion, the dialogue explanation would help them identify the discrepancy between the facts and rules they had been given, compared to the facts and rules the computational system was using. Each participant was presented with two scenarios out of a possible six, and 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 tree

[2] Full details of this process with proofs that dialogues terminate and identify a difference in user facts and rules where one exists can be found at https://github.com/louiseadennis/bluebook/blob/main/bluebook5.pdf

explanation, and 18 (75%) found the dialogue explanation "Easy" (see Fig. 3). The results also show that dialogue explanation is helpful for finding the dis-



(a) (b)

Fig. 3: Dialogue Explanation Results on how difficult the users found the dialogue system (a), and whether the explanation was helpful (b)

agreement. In this experiment, we consider only proof trees with a depth of fewer than ten levels. In the future, we might more complex situations including more extensive information or rules.

4 Conclusion

We have proposed a dialogue explanation approach to explain the reasoning in systems where derivations are represented as trees, typical of rule-based AI systems. A dialogue system assumes that an explanation is a collaborative process in which the system determines what information it is that the user wants. A dialogue explanation allows the user and system to co-create an explanation based on the user's context which is preferable to a graph representing the full dialogue in most cases. The result of the user study with 24 volunteers also shows that the majority prefers our dialogue system to a tree representation of the computer's decision-making process. This viable mechanism empowers machines with the human ability to explain their actions and significantly furthers our knowledge of the conversational approach to explainable AI.

In future work we hope to examine how our explanations could be adapted to explore "what-if" scenarios which would allow a dialogue to progress beyond identifying a source of disagreement to exploring whether eliminating that disagreement would change the system's conclusion.

References

1. Arioua, A., Tamani, N., Croitoru, M.: Query answering explanation in inconsistent datalog+/- knowledge bases. In: Database and Expert Systems Applications. pp. 203–219. Springer (2015)

2. Brarda, M.B., Tamargo, L.H., García, A.J.: An approach to enhance argument-based multi-criteria decision systems with conditional preferences and explainable answers. Expert Systems with Applications **126**, 171–186 (2019)
3. Briguez, C.E., Budan, M.C., Deagustini, C.A., Maguitman, A.G., Capobianco, M., Simari, G.R.: Argument-based mixed recommenders and their application to movie suggestion. Expert Systems with Applications **41**(14), 6467–6482 (2014)
4. Cocarascu, O., Stylianou, A., Čyras, K., Toni, F.: Data-empowered argumentation for dialectically explainable predictions. In: ECAI 2020, pp. 2449–2456. IOS Press (2020)
5. Dennis, L.A., Oren, N.: 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) (2021)
6. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
7. Johnson-Laird, P.N.: Mental models in cognitive science. Cognitive science **4**(1), 71–115 (1980)
8. Lacave, C., Diez, F.J.: A review of explanation methods for heuristic expert systems. The Knowledge Engineering Review **19**(2), 133–146 (2004)
9. Lin, B.Y., Chen, X., Chen, J., Ren, X.: Kagnet: Knowledge-aware graph networks for commonsense reasoning. arXiv preprint arXiv:1909.02151 (2019)
10. Oren, N., Deemter, K.v., Vasconcelos, W.W.: Argument-based plan explanation. In: Knowledge Engineering Tools and Techniques for AI Planning, pp. 173–188. Springer (2020)
11. Reggia, J.A., Perricone, B.T.: Answer justification in medical decision support systems based on bayesian classification. Computers in Biology and Medicine **15**(4), 161–167 (1985)
12. Rodríguez, P., Heras, S., Palanca, J., Poveda, J.M., Duque, N., Julián, V.: An educational recommender system based on argumentation theory. AI Communications **30**(1), 19–36 (2017)
13. Rolf, L., Kern-Isberner, G., Brewka, G.: Argumentation-based explanations for answer sets using adf. In: International Conference on Logic Programming and Nonmonotonic Reasoning. pp. 89–102. Springer (2019)
14. Sendi, N., Abchiche-Mimouni, N., Zehraoui, F.: A new transparent ensemble method based on deep learning. Procedia Computer Science **159**, 271–280 (2019)
15. Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., Cohen, S.N.: An artificial intelligence program to advise physicians regarding antimicrobial therapy. Computers and Biomedical Research **6**(6), 544–560 (1973)
16. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data & knowledge engineering **25**(1-2), 161–197 (1998)
17. Swartout, W.R.: Xplain: A system for creating and explaining expert consulting programs. Artificial intelligence **21**(3), 285–325 (1983)
18. Timmer, S.T., Meyer, J.J.C., Prakken, H., Renooij, S., Verheij, B.: A two-phase method for extracting explanatory arguments from bayesian networks. International Journal of Approximate Reasoning **80**, 475–494 (2017)
19. Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and explainable artificial intelligence: a survey. The Knowledge Engineering Review **36** (2021)