Dialogue Explanation With Reasoning For AI

Yifan Xu Supervised by Joe Collenette, Louise A. Dennis, Clare Dixon

Department of Computer Science, The University of Manchester

{yifan.xu,joe.collenette,louise.dennis,clare.dixon}@manchester.ac.uk

The University of Manchester

Overview

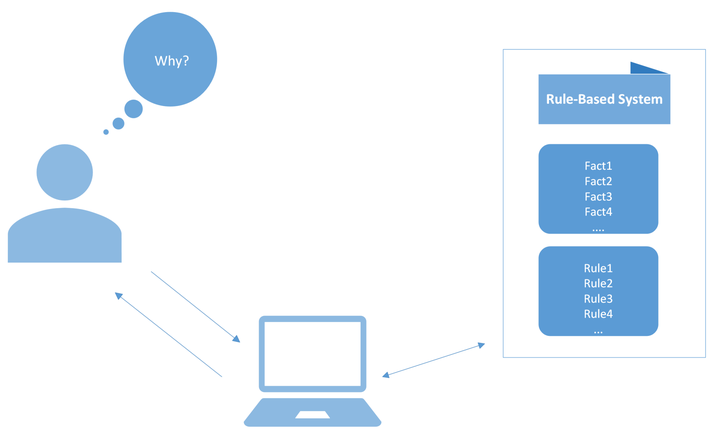
Our hypothesis is that when a system makes a deduction that was, in some way, unexpected by the user then locating the source of the disagreement or misunderstanding is best achieved through a collaborative dialogue process that allows the participants to gradually isolate the cause.

- **Q1** *Can dialogue provide an understandable explanation for rules-based reasoning?*
- **Q2** *Can dialogue explanation provide an understandable explanation for an AI system with learned rules?*
- We measure understandability by how easy it is for a user to locate the cause of a disagreement between themselves and the system.

Rules, Facts and Deductions

Covid Advice system

- It is a rule-based system consists of a set of initial facts, F, of positive literals in $\mathcal{L}$; and a set of rules, R.

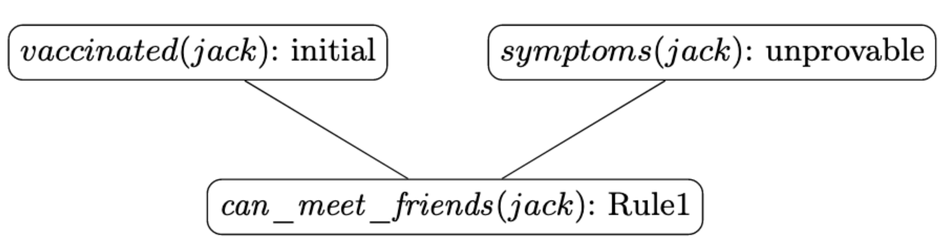


Covid Advice system

- Its goal is to provide users with an **one-step** explanation for any particular why or why not questions about Covid rules and regulations.
- **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 $l \in L \backslash \{initial, unprovable\}$.
- **A Fact** is a statement that the system either knows at the start of reasoning (provided as part of an initial problem statement) or have been deduced during the course of reasoning

A Directed Acyclic Graph

– **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.



A Proof Tree

Dialogue Mechanism

There are six possible statements that can be made in the course of a dialogue:

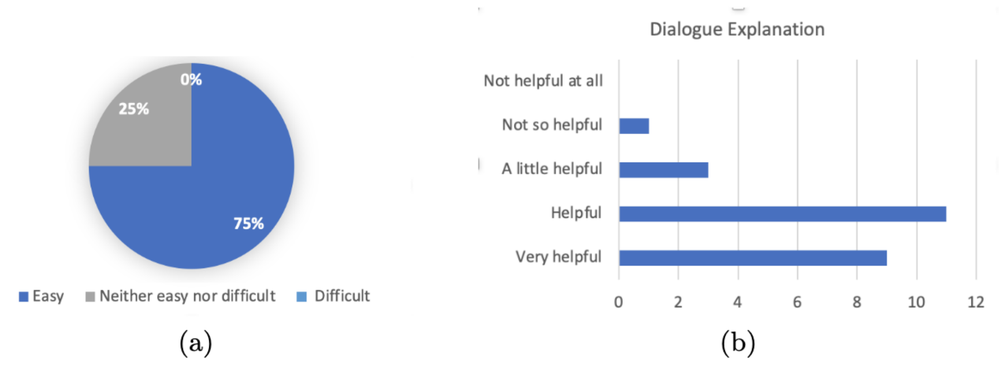
1. $different\_fact(t, i, j)$ – i has t as an initial fact and j does not.
2. $different\_rule(l : A \rightarrow C, i, j)$ – i has $l : A \rightarrow C$ as a rule and j does not.
3. $initial(t)$ – t is an initial fact for the Player.
4. $l : a \rightarrow t$ – the player deduced t from the terms in a using the rule labelled l
5. $why(t)$ - why do you believe t?
6. $whynot(t)$ – why don't you believe t?

Computer Rule: 1. If A has pinged, and A is vaccinated, and A has not taken a test, then A need 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: Because 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!

Dialogue explanation example where the user and computer disagree

User Evaluation

- 83.3% preferred dialogue explanation to the tree explanation
- 18 (75%) found the dialogue explanation easy to understand



(a) (b)

Dialogue explanation

Future work

- Develop a Neuro Symbolic AI system with a dialogue mechanism, and conduct a user evaluation for such a system.
- A neural network-based advice system or open-source training data set (e.g., for medical diagnosis), then extract from it a rule-based system using the REX methodology [1] and it is that rule-based system that will then offer advice.

References

[1] Zohreh Shams, Botty Dimanov, Sumaiyah Kola, Nikola Simidjievski, Helena Andres Terre, Paul Scherer, Urska Matjasec, Jean Abraham, Mateja Jamnik, and Pietro Liò. Rem: An integrative rule extraction methodology for explainable data analysis in healthcare. *medRxiv*, 2021.

[2] Yifan Xu, Joe Collenette, Louise A. Dennis, and Clare Dixon. Dialogue explanation with reasoning for ai. In *Explainable Logic-Based Knowledge Representation (To Appear)*, 2022.