User Evaluated Explainable BDI Agents

Joost Broekens
Maaike Harbers
Koen Hindriks
Karel van den Bosch
Catholijn Jonker
John-Jules Meyer

Background

- Explainable autonomous agents...why?
  - VR Training
  - Games
  - Interactive storytelling
    - Flexibility and believability
  - Agent debugging
    - Use agent reasoning-language translation
- Related to expert system, but not the same.
  - Main diff: expl. for *actions* derived from autonomous behavior and reasoning, not necessarily a conclusion of a system build for explanation

Background

- Main aim:
  - explain autonomous actions in for humans understandible way.
- Two main streams
  - Directly derived from agent behavior and reasoning

    Benefit: Agent model is all, Drawback: quality of explanation?
  - Additional information and or knowlegde is needed
    - Benefit: fine tune explanation, Drawback: more modeling

- Our appraoch:
  - BDI Agent, directly derived.

Main question

- Is it possible to generate useful explanations for an agent's actions based only on (the history of) that agent's BDI state?
- How does the usefulness depend on the type of action to be explained?
- Are there guidelines on how to do it?
  - For explainable BDI Agent development
  - For generating the explanation

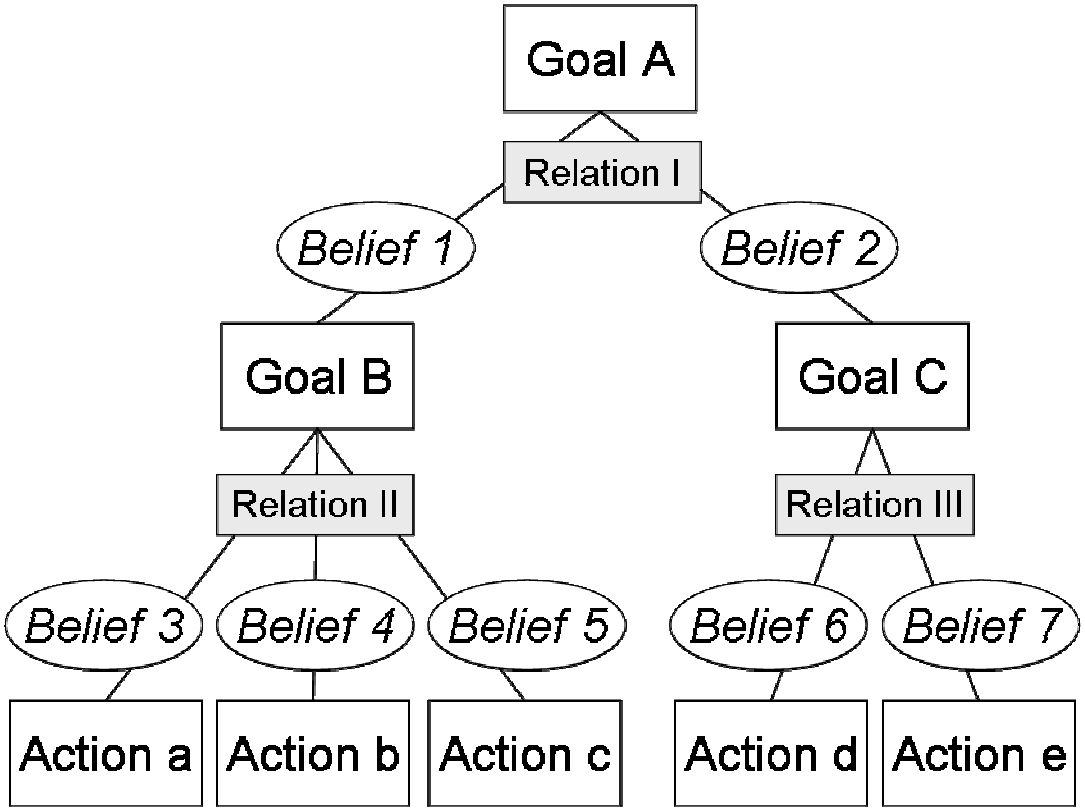
"Loose" hypothesis

- Given a BDI agent:
- Given an explanation mechanism:
- Given different types of action

- The usefulness of an explanation mechanism depends on the type of action to be explained generated by the BDI agent.

Action types

- All: *and* relation between siblings
- Seq: *and* relation in order
- One: *xor* relation between siblings

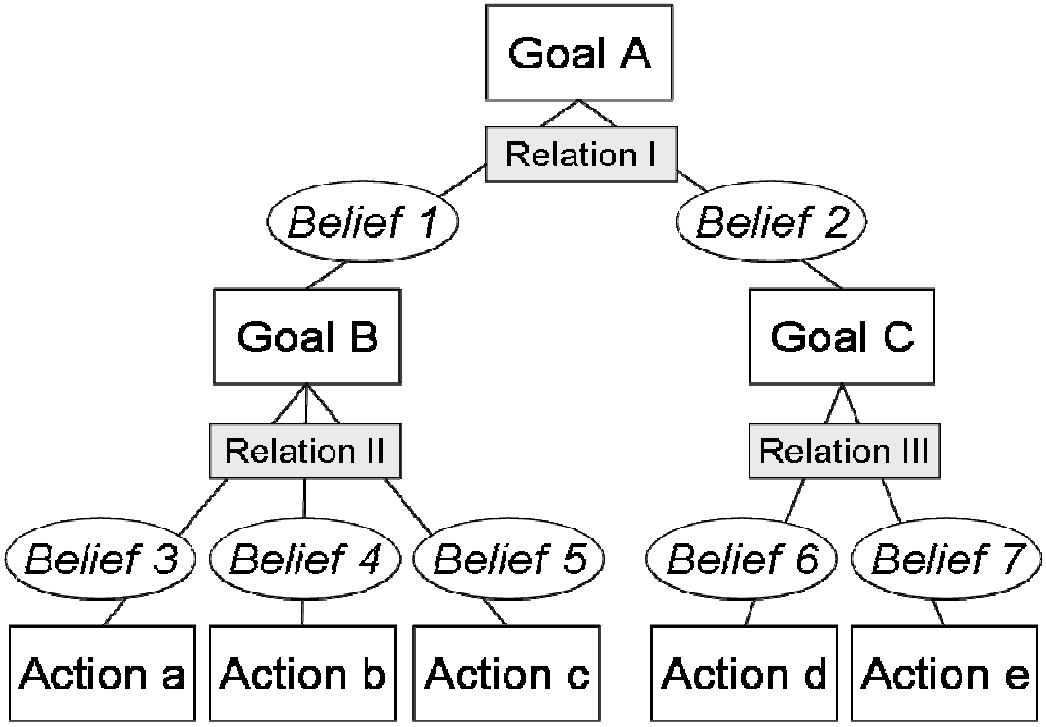


BDI Agent


1 I want to eat pancakes
2 I have not collected everything I need
12 I have collected all necessaries
28 The pancake mix is ready
3 I want to collect everything I need
13 I want to make pancake mix
29 I want to bake pancakes
4 I have not collected ingredients
10 I have not collected kitchen stuff
14 Not all ingredients are in bowl
22 All ingredients are in bowl
30 The frying pan is not heated
32 The frying pan is heated
34 The butter is melted
36 The first side is ready
42 The pancake is ready
5 I want to collect all ingredients
11 2. I take the required kitchen stuff
15 I want all ingredients in the bowl
23 I want to mix the ingredients
31 7. I heat the fry-ing pan
33 8. I put butter in the frying pan
35 9. I put the mix in the frying pan
37 I want to flip the pancake
43 11. I put the pancake on a plate
6 Not all ingredients are home
8 All ingredients are home
16 I have not added eggs
18 I have not added flower
20 I have not added milk
24 I have a mixer
26 I have no mixer
38 I have no spatula
40 I have a spatula
7 I buy the ingredients
9 1. I take the ingredients from the cupboard
17 3. I put eggs in the bowl
19 4. I put flower in the bowl
21 5. I put milk in the bowl
25 6. I mix the ingredients with a mixer
27 I mix the ingredients by hand
39 10. I flip the pancake by throwing it
41 I flip the pancake with a spatula

Explanation mechanisms

- Alg 1: Explain action using nearest goal up
- Alg 2: Explain action using enabling belief
- Alg 3: Explain action using next goal/action to be achieved/executed



Detailed hypothesis

- The usefulness and naturalness of an explanation generated from a resulting goal-plan tree of a BDI agent depends on
  - the interaction between
    - action type (all, seq, one) and
    - explanation mechanism (goal up, enabling believe, next goal/action).

Experiment

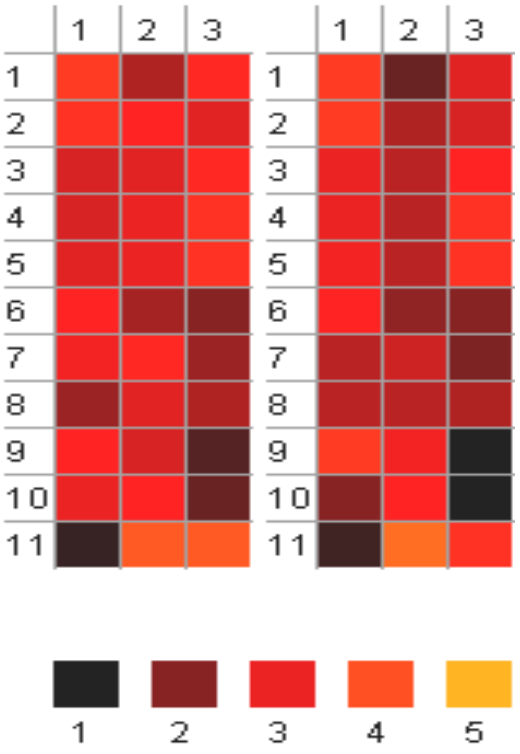
- Simple domain: cooking pancakes
- 30 users, 10 per explanation mechanism.
- 11 actions of different types.
- Each mechanism generates explanation for each of the 11 actions
- Each user rated 11 actions.
- Each user gave its own explanation based on available elements in goal-plan tree.

Example questionaire

- (Show questionaire for condition 1)

Rating results

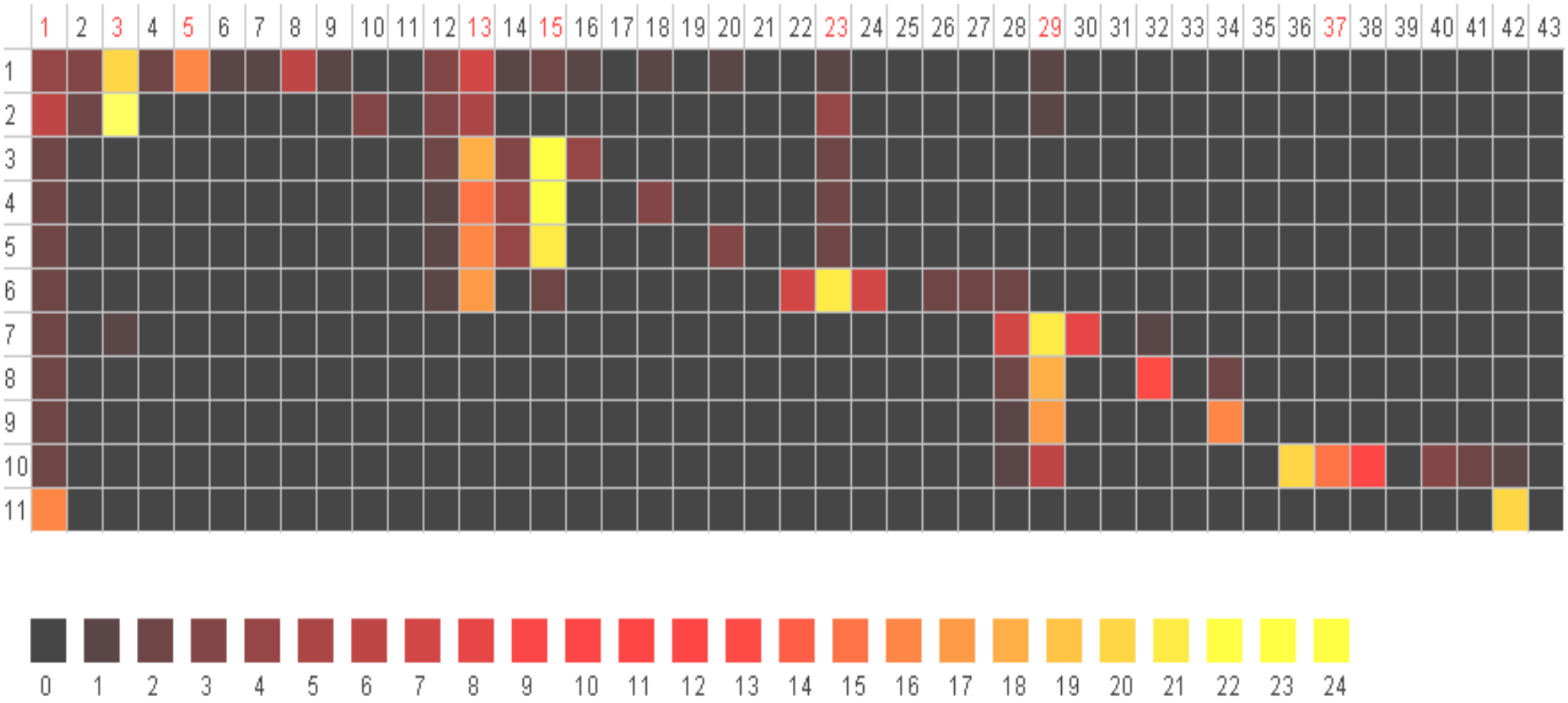
- Analysis showed two important things:
  - Explanations with parent goal are slightly better on average
  - Important interaction effect between action and expl, but not as hypothesized.
    - e.g. a*ll* actions 3,4,5 better explained using next goal



- In detail:
  - left=usefulness, right=naturalness
  - x=algorithm
  - y=action

Users own choices results

- Based on tree feedback



Concluding observations

- Main hypothesis confirmed
- Not as straighforward as we hypothesized
- All actions seem to need at least one extra element in addition to parent goal.
  - e.g. type one (or):
    - Enabling condition, parent's condition, and parent goal
  - e.g. for type seq:
    - Enabling condition needed

Hypothesized guidelines

- Parent goal is most important default, make it meaningful
- XOR /or choices should be modelled with meaningful subgoal (not abstract choice action) or use more elements.
- Actions starting new phase: include parent goal and enabling condition of parent goal.
- Actions in sequence: include enabling condition.