Explaining Aggregate Behaviour in Cognitive Agent Simulations using Explanation

Tobias Ahlbrecht, Michael Winikoff

14 May 2019 @ EXTRAAMAS

The story so far

This morning:

- **cognitive** agents
- explanations
  - reasons = **b**eliefs, **d**esires, **v**aluings, ...
- existing work: focus on **single** agents

What now?

Goal

- generate explanations
  - for **multiple** agents
  - in **simulation** context

Approach:

- use existing method
- apply to **groups** of agents

Framework

Simulation setting:

- **BDI** agents (“Jason-style”) + environment
- no **coordination**

Explanation:

- Simplification (for now) of
  - Winikoff, M., Dignum, V., Dignum, F.: Why bad coffee? explaining agent plans with valuings. In: Gallina, B., Skavhaug, A., Schoitsch, E., Bitsch, F. (eds.) Computer Safety, Reliability, and Security. pp. 521–534. LNCS 11094, Springer (2018)

Explaining

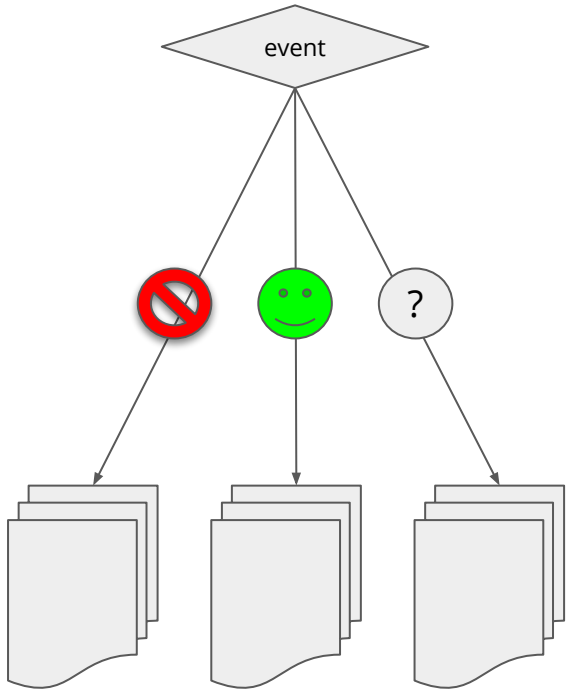
Question: “**Why** did the agent perform **action** A?”

Input:

- (actual) trace of actions
- agent program (as goal-plan tree)
- query (i.e. an action)

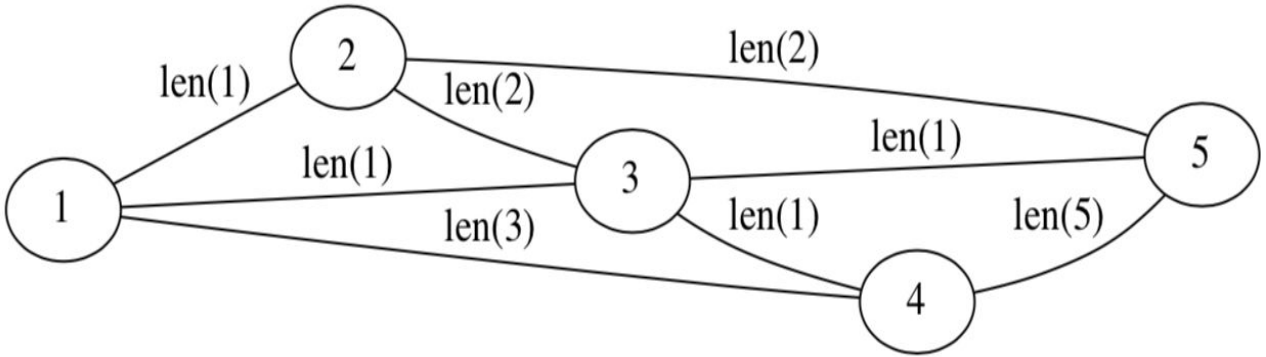
Explanation mechanism

- cut off trace after the action
- collect **explanatory factors**
  - action preconditions
  - context that enabled the action



Simple example: Traffic

- graph of **streets** and **intersections**
- road **preferences**
  - traffic
  - length
  - bridges



Example continued

Actions:

- `.takeRoad(Road)`
- `.getDetour(Road)`
- `.getTraffic(Road)`
- …

Plans:

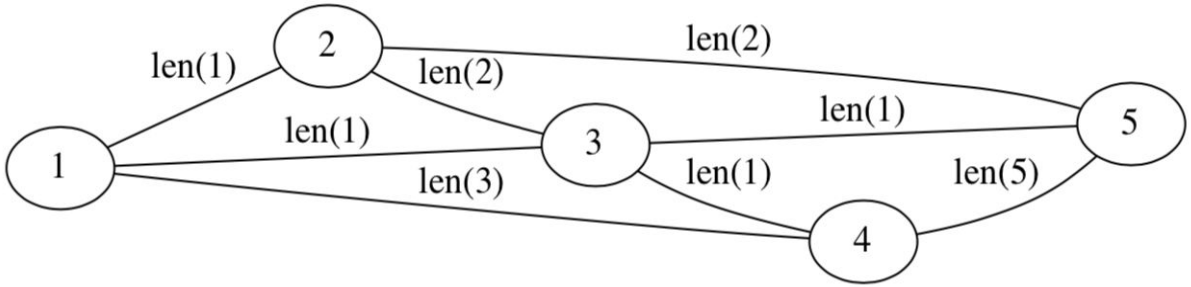
```
+!goto(To) : position(Pos) ∧ bridge(Pos, To) ∧
.bridgeStatus(Pos, To, open(false)) ∧ not plannedRoute(_)
   ← !useDetour(To).
```

Example explanation

Car from 1 to 5

Why road (1,2)?

```
reach(5)
notAtDestination & noPlannedRoute
would_prefer_due_to_traffic([1, 2], [1, 3])
would_prefer_due_to_route_length([1, 2], [1, 4])
goto(2)
noBridge(1, 2)
```

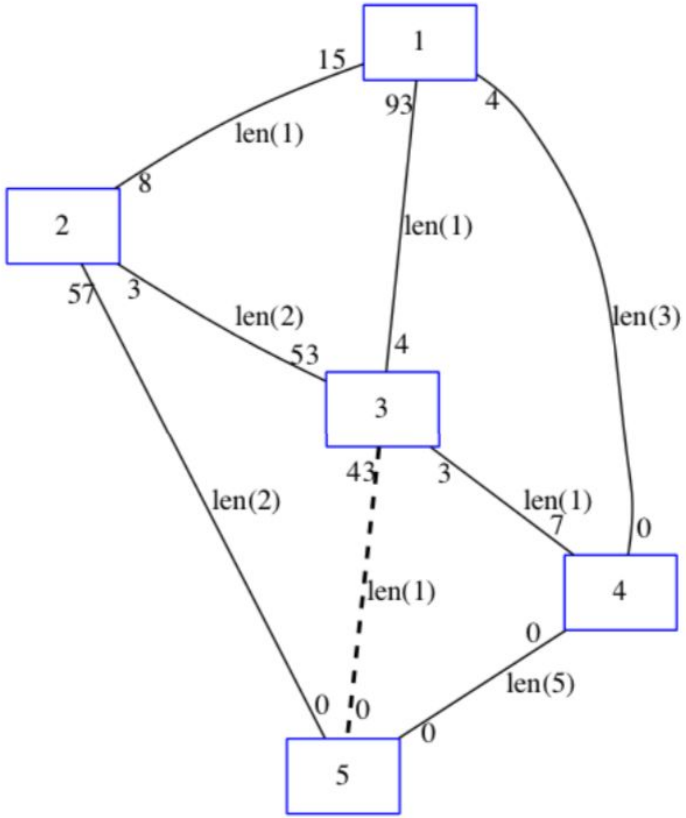


Aggregating explanations

“Why did the agent**s** perform action A?”

- identify relevant agents
- generate explanations
- **count** occurence of each explanatory factor

Sample



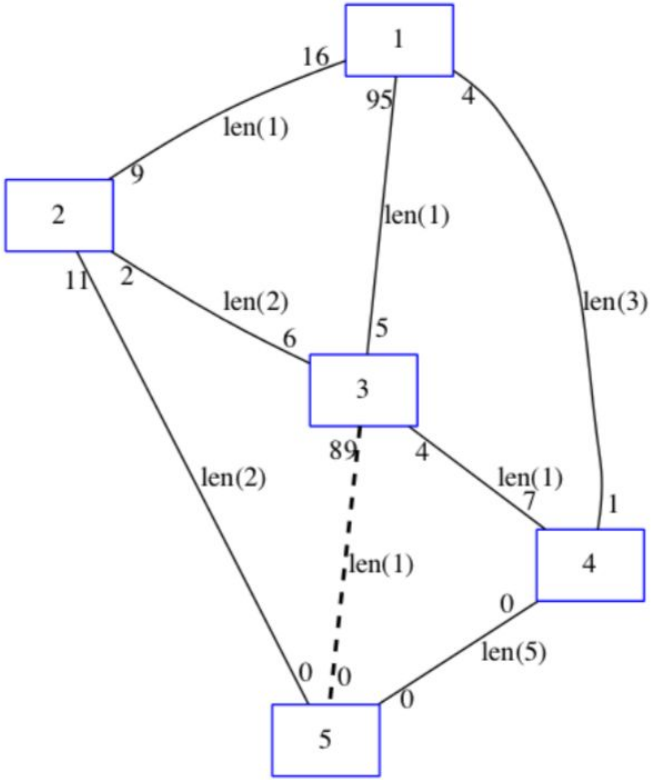
Why use road (2,5)?

- 57 times noPlannedRoute
- 57 times noBridge(2, 5)
- 57 times notAtDestination
- 57 times reach(5)
- 53 times would_prefer_due_to_route_length([1, 3], [1, 4])
- 53 times noBridge(3, 2)
- 53 times would_prefer_due_to_route_length([3, 5], [3, 4])
- 53 times notWaitForClosedBridge(3, 5)
- 53 times would_prefer_due_to_route_length([3, 5], [3, 2])
- 53 times notAtDestination(2)
- 53 times useDetourAround(3, 5)
- 53 times would_prefer_due_to_route_length([3, 5], [3, 1])
- 53 times plannedRoute([5|[]])
- [...]

Test hypothesis

Rerun:

(bridge less often closed)



Process

- locate interesting/surprising behaviour
  - pose a question
- generate an answer
- follow-up questions
- counterfactual experiments

Result

- use **explanations** to understand a simulation
- **test hypothesis** with reruns
- identify **“odd” simulation behaviour** (not pictured)
  - happened to us

Future work - Open Challenges

Specific:

- human participant evaluation
- presentation
- improve aggregation
- leverage environment knowledge

General:

- explanations for jointly-acting agents (**teams**)
- **integration** to MAS explanation

Thank you!

Questions?

Nothing to see here

More query types

“Why did the agents perform action A under **condition** C?”

“Why are agents in **situation** S?”