Policy Graphs and Intention

Answering *why* and *how* from a telic perspective

Victor Gimenez-Abalos*,
Sergio Alvarez-Napagao*,
Adrian Tormos, Ulises Cortés,
Javier Vázquez-Salceda

Motivation and context

- Agents, as an AI system, are relevant targets for explainability: XAI
- But! Some explanations are better than others
- e.g. according to Herbert Grice they can be rated by...
  - *manner* (how interpretable for the receiver)
  - *quality* (truthfulness)
  - *quantity* (conciseness)
  - *relation* (relevance)

Motivation and context

- Agents, as an AI system, are relevant targets for explainability: XAI
- But! Some explanations are better than others
- e.g. according to Herbert Grice they can be rated by...
  - ***manner* (how interpretable for the receiver)**
  - ***quality* (truthfulness)**
  - *quantity* (conciseness)
  - *relation* (relevance)

Motivation and context

reliability vs **interpretability**

is the explanation factually correct?

depends on the sender

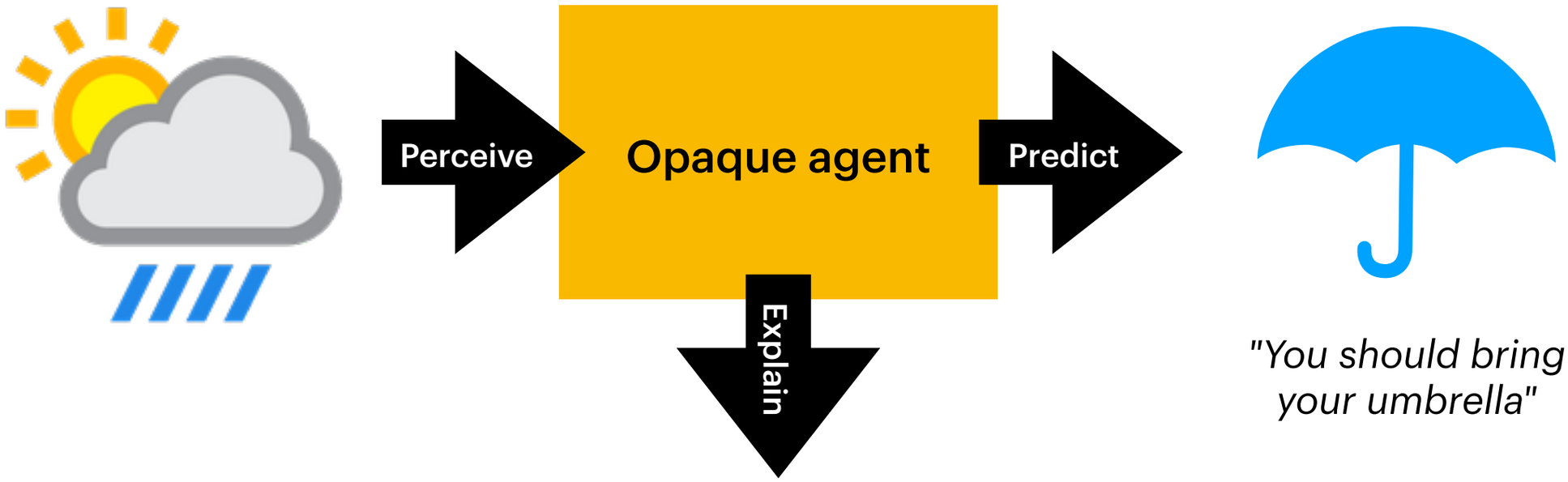
CONFLICT/TRADE-OFF

how much of the explained behaviour can the receiver comprehend?

depends on the receiver

What is explainability used for? Justify / Control / Improve / Discover

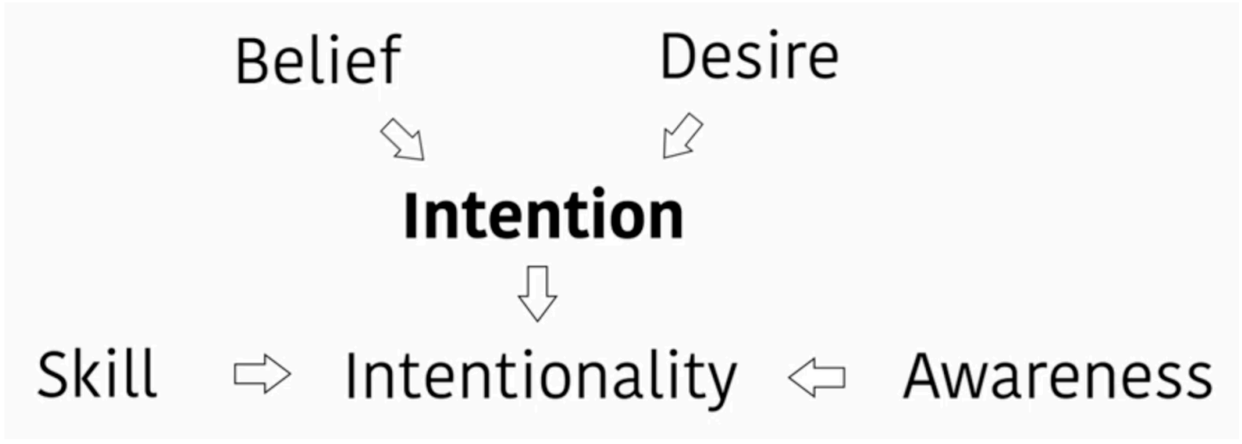
Motivation and context



- A human explainee would likely be *happy* with this explanation
  - Vulnerability against cognitive biases: confabulation, antropomorphisation
- There is nothing in this procedure that allows us to assess the reliability of this explanation
  - e.g. the dataset could be biased towards bringing an umbrella in sunny days

Assessing truthfulness

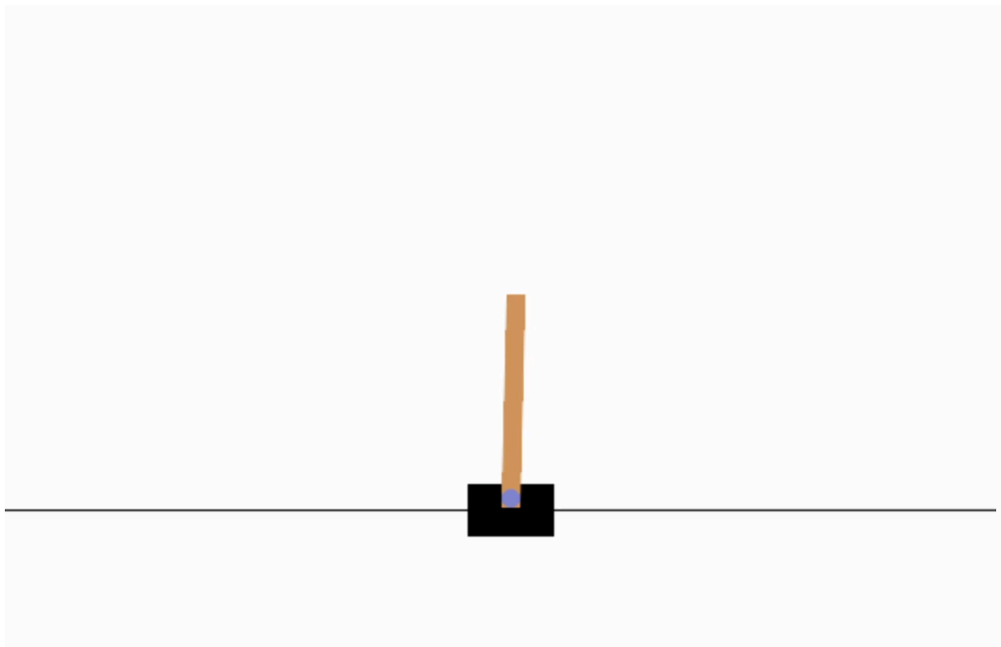
- What possible metrics can there be for evaluating the truthfulness of explanations (*in the context of agent behaviour*)?
- [Malle, 1997/2022]: humans tend to provide intentional explanations for behaviours (under certain conditions)

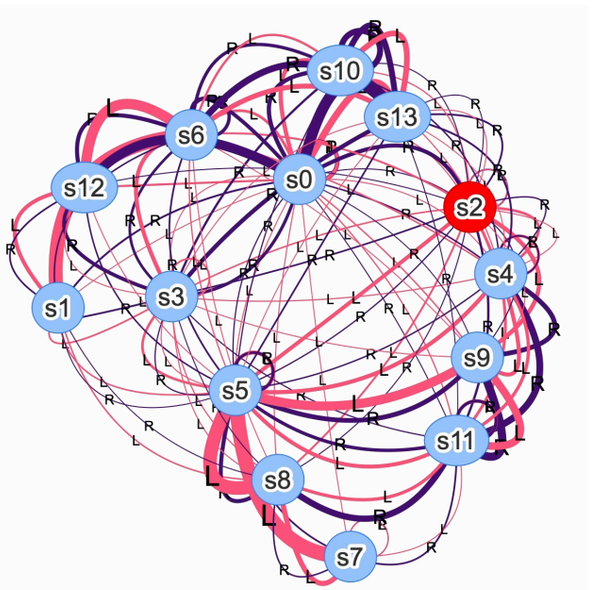


- Intentional explanations are causal-based
  - But the causes of an intention can be (and often are) about future states of the world

Policy graphs

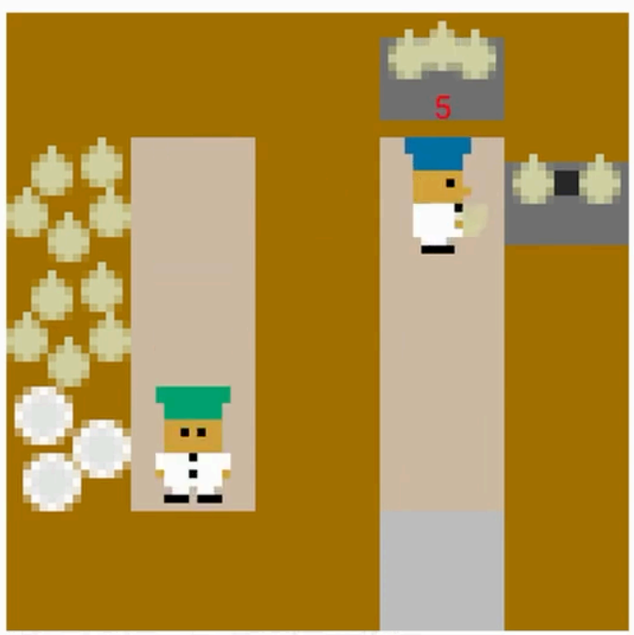
- Formalism for capturing the behaviour of the agent in the environment
  - Frequentist estimation of World model+Agent policy: $p(s' \mid a, s) \cdot \pi(a \mid s)$
  - Based on discrete representations of state and actions
- Allows to answer short-sighted questions, e.g. *What would you do in s?*

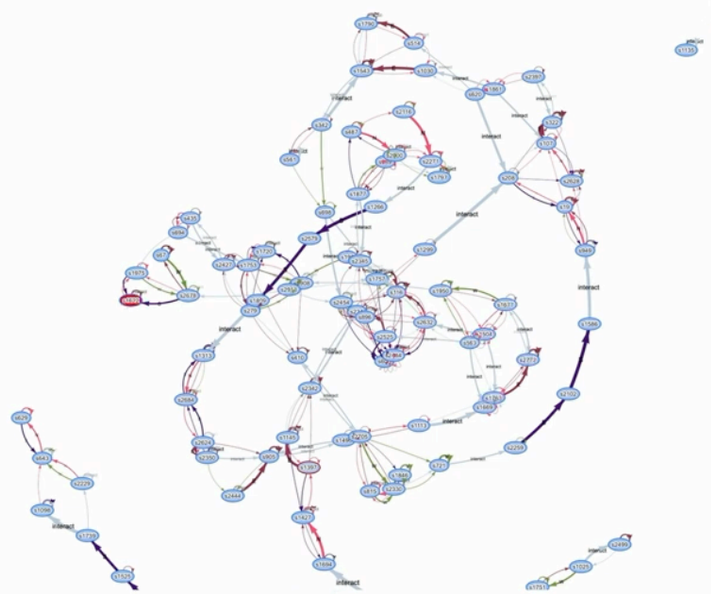




Policy graphs

- Formalism for capturing the behaviour of the agent in the environment
  - Frequentist estimation of World model+Agent policy: $p(s' \mid a, s) \cdot \pi(a \mid s)$
  - Based on discrete representations of state and actions
- Allows to answer short-sighted questions, e.g. *What would you do in s?*





Intentional Policy Graphs

- We present Intention-aware Policy Graphs, extending the original model with inferred intentions
- Intention metrics as probabilities, based on:
  - The possible trajectories of agents (from generated policy graphs)
  - User-defined desires
- The explainee makes hypotheses on the agent's desires
- Intentions are then the probability that a desire is fulfilled

$$P(s \in S(I_d))$$

attributed intention probability

$$\mathbb{E}_{s \in S(I_d)}\left(I_d(s)\right) = \sum_{s \in S(I_d)} I_d(s) * \frac{P(s)}{P(s \in S(I_d))}$$

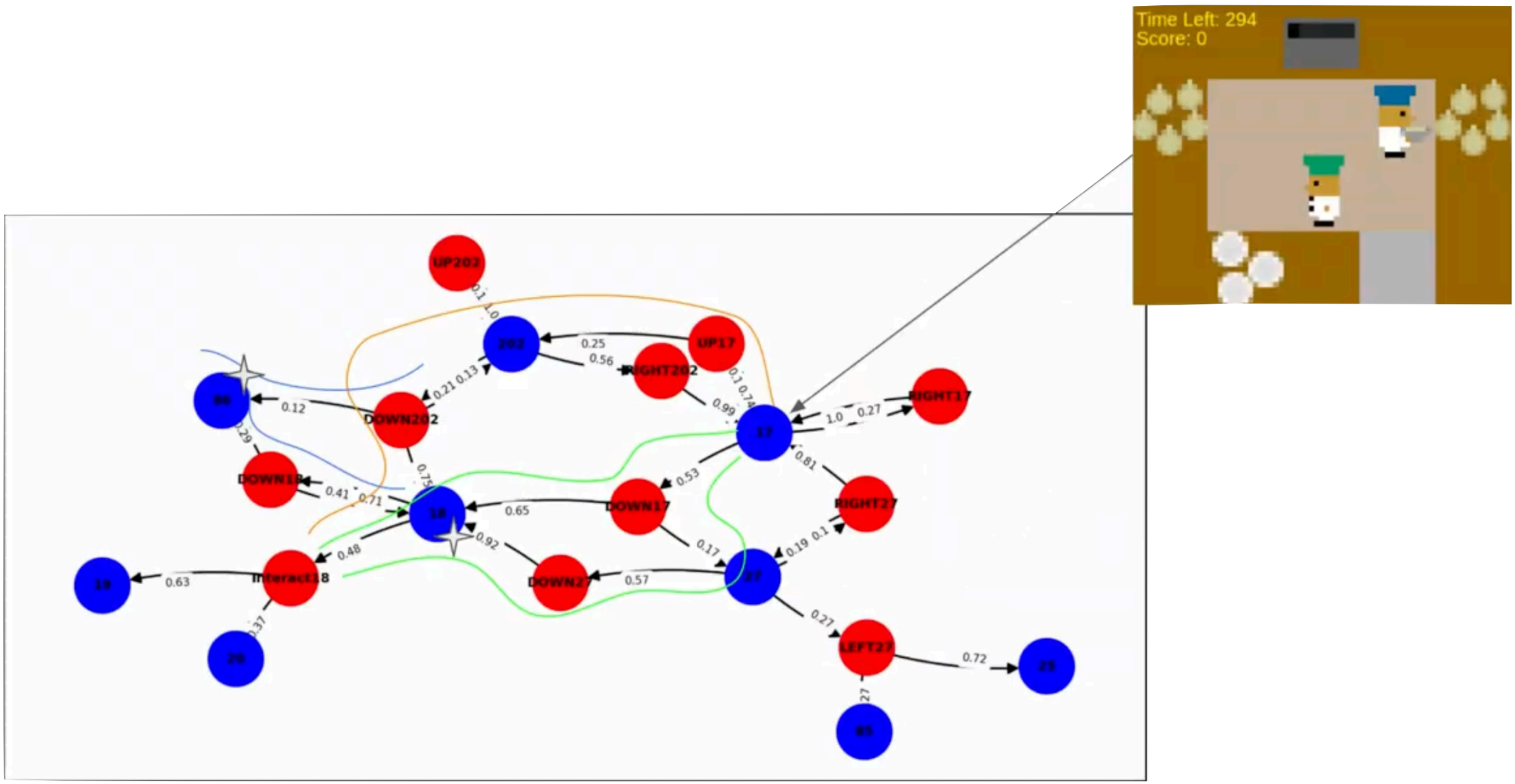
expected intention probability

We also define a *commitment threshold* C that allows to tradeoff reliability and interpretability:

$$I_d(s) \geq C$$

The agent is said to have (at least some) intention to fulfill d

Metrics for intention



What can we explain with IPGs?

- *What* do you intend to do in state s?

$$\{d \mid I_d(s) \geq C\}$$

- *Why* do you do action a in state s?

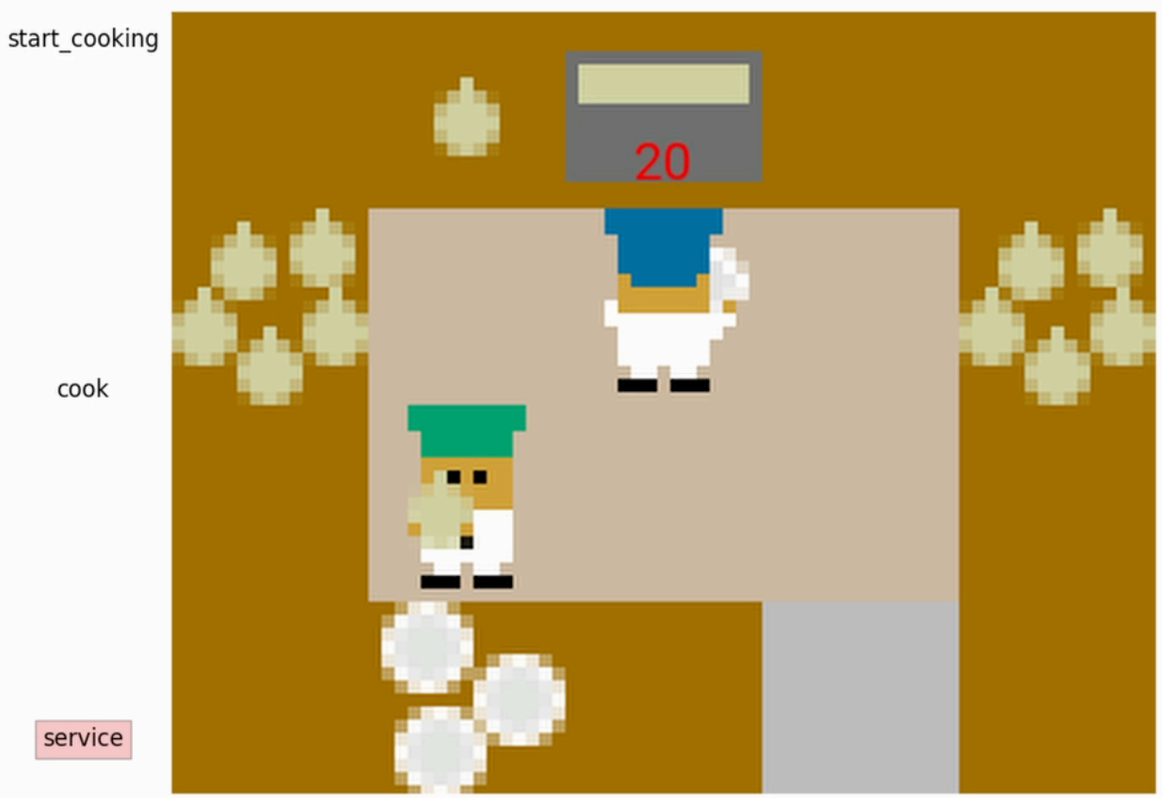
$$\mathbb{E}_{P(s'|a,s)}(I_d(s')) - I_d(s) = \sum_{s'} P(s' \mid a, s) * I_d(s') - I_d(s)$$

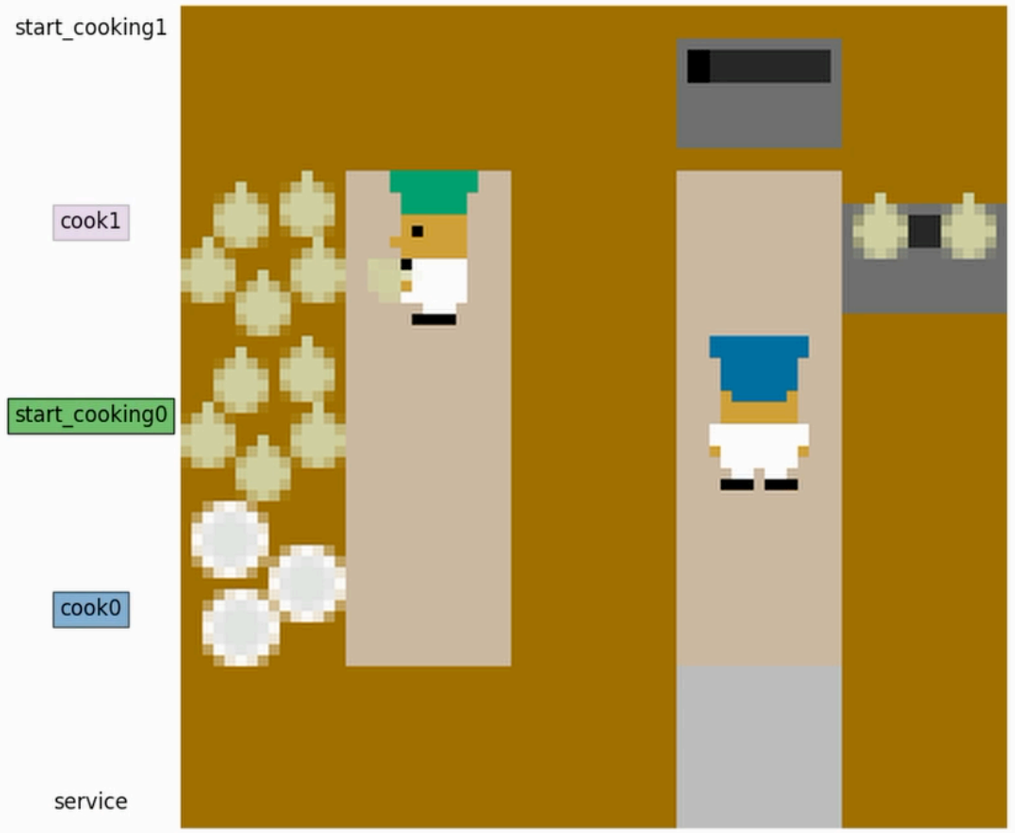
- *How* do you plan to fulfill intention I_d from s?

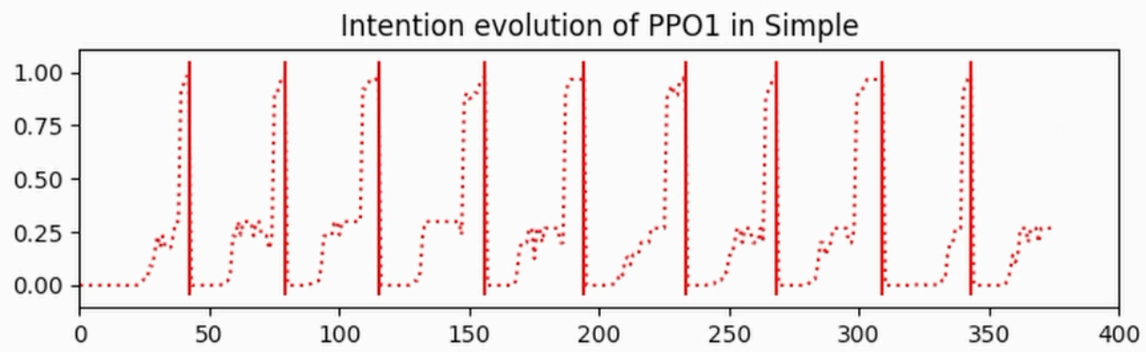
procedure HOW(d, s, PG)
  $current \leftarrow s$
  if $s \vdash d$ **then** ▷ State can fulfill desire
    return a_d ▷ return action that fulfills the desire
  end if
  $s' \leftarrow argmax_{s', a \in Succ(s)} I_d(s')$
  return Concat(a, s', how(d, s', PG))
end procedure

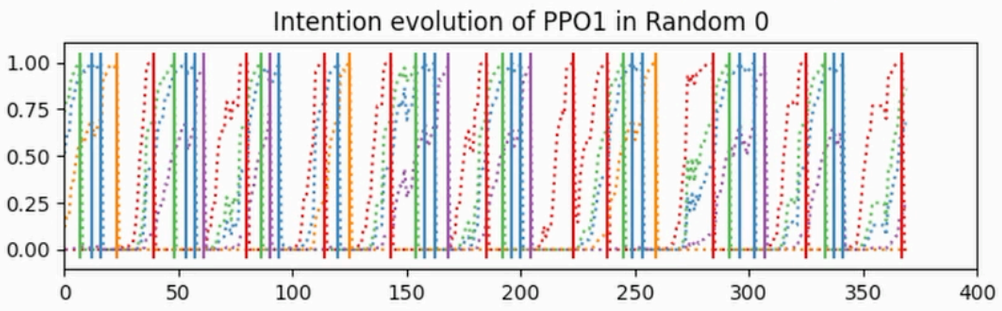
Example: Overcooked-AI

Intention evolution



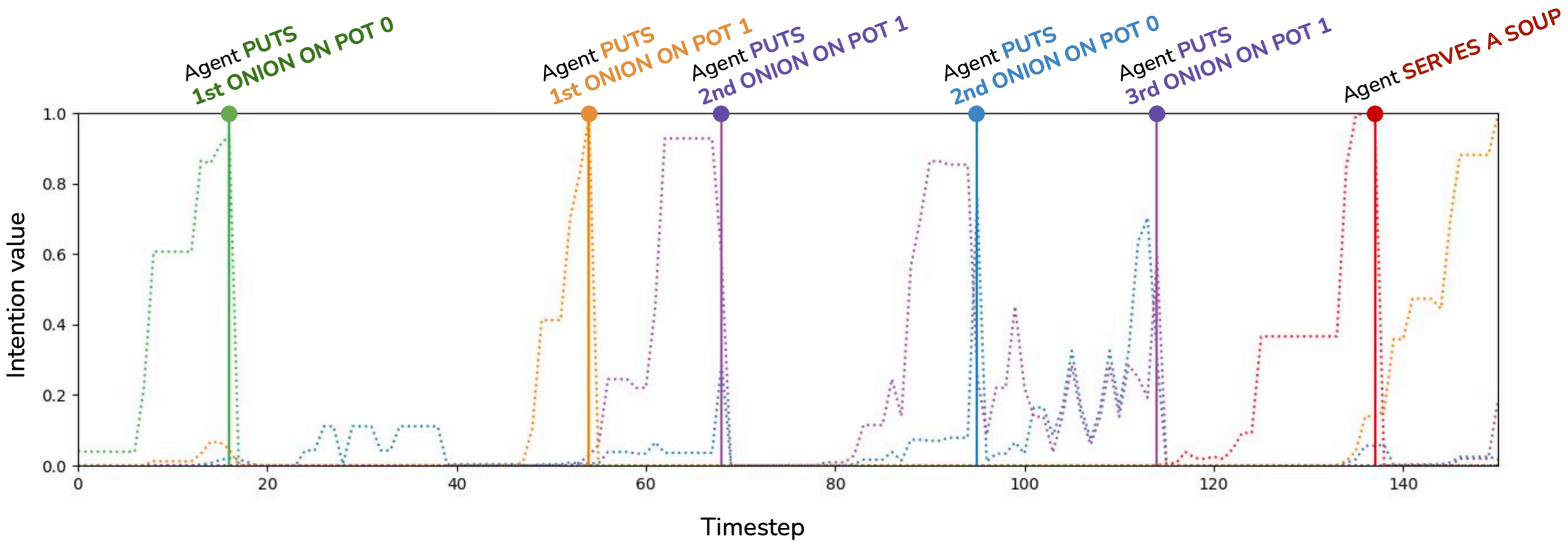






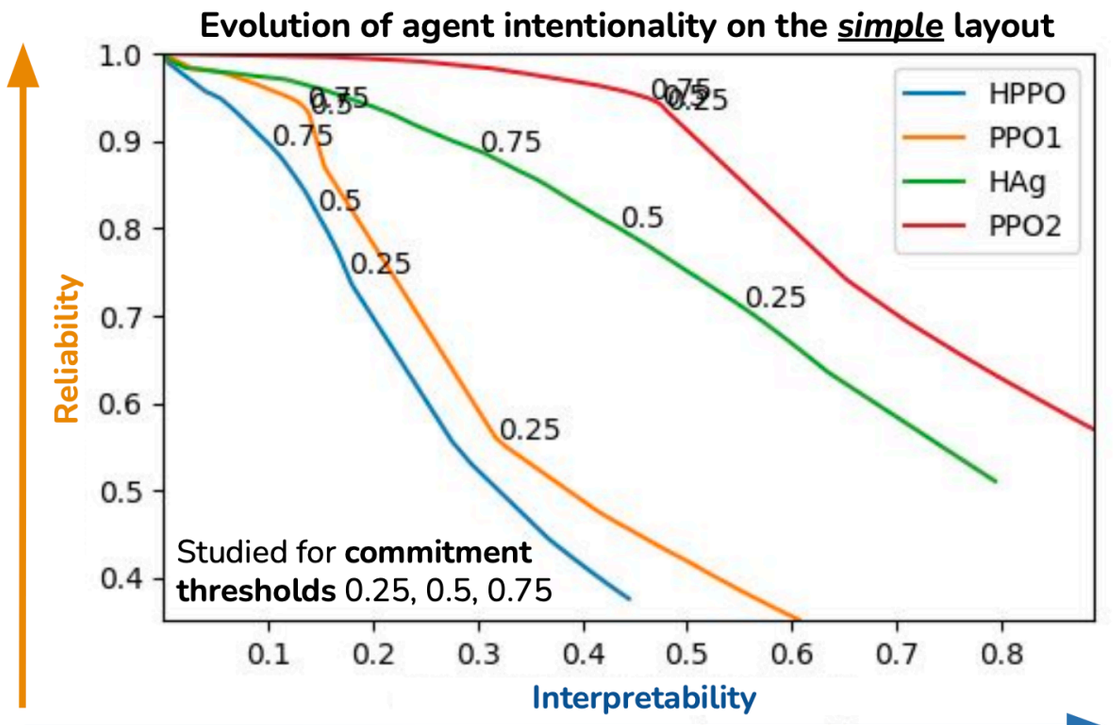
Example: Overcooked-AI

Intention evolution



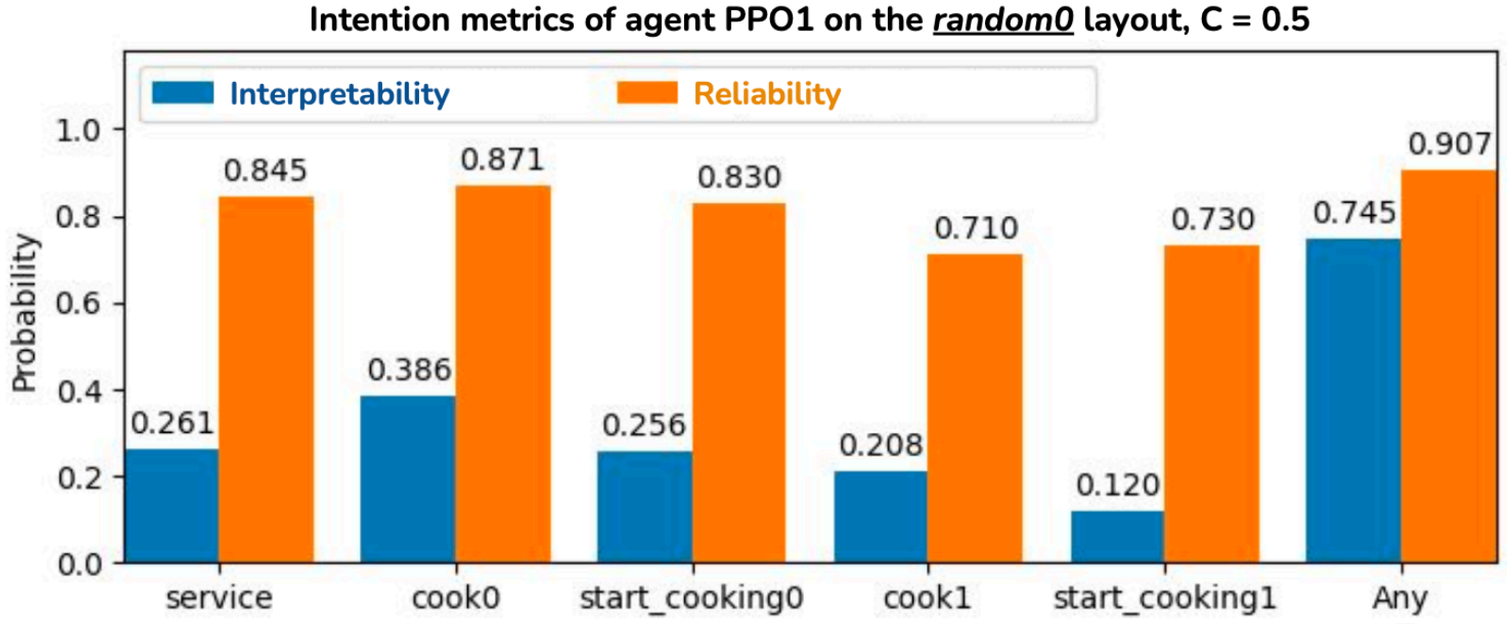
Example: Overcooked-AI

Intention metrics



Reliability: How often does the attributed intention end up happening?

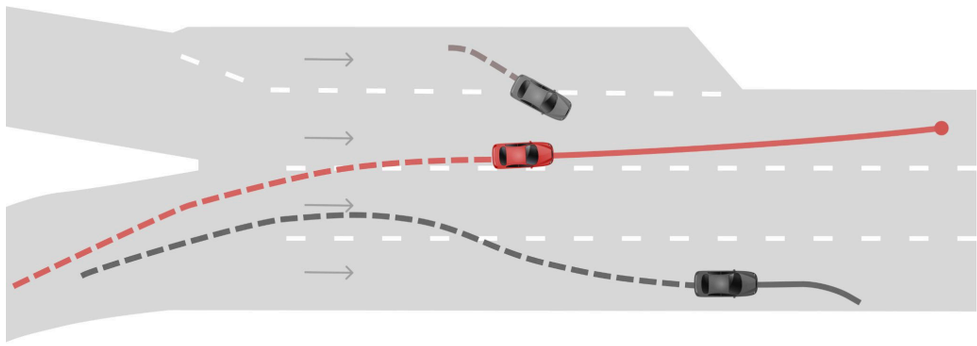
Interpretability: How often is the intention attributed? For how long is the agent in a state that can be explained?



Conclusions

- We are able to attribute desires and intentions to opaque agents
- The explainee can ask for explanations of those behaviours that seem reasonable to them
- We introduce a threshold to adjust the trade-off between reliability and interpretability
- This formalism allows us to answer several questions that are *telic* in nature
- Don't enjoy cooking scenarios? Check our latest work on applying IPs into autonomous vehicles scenarios

IPGs in AVs (EXTRAAMAS)



Thanks for attending!

Any questions?

{victor.gimenez, sergio.alvarez}@bsc.es

This paper

IPGs in AVs (EXTRAAMAS)

More on intentions (EXTRAAMAS)