Institut de Robòtica
i Informàtica Industrial

CSIC

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

PhDday 2024

Fostering human-robot mutual understanding by explaining the internal beliefs

Tamlin Love

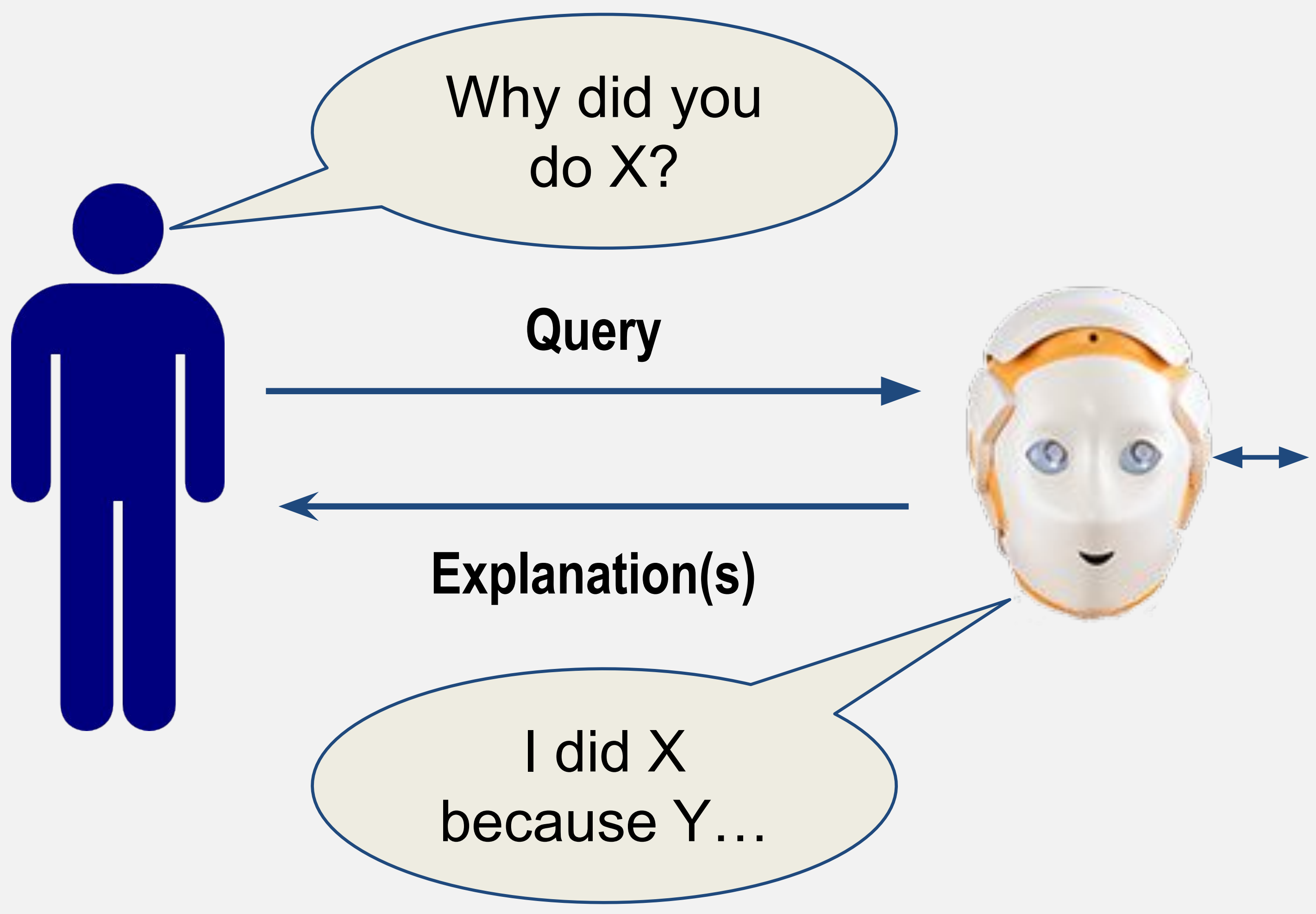
Supervisor: **Guillem Alenyà Ribas**

MOTIVATION

- Robots are more and more involved in our daily lives
- In order to **improve trust** and **collaboration** between humans and robots, we need to **understand** how they make decisions
- **Explainability** can help with this…
- …but explainability is still a **challenge**, especially for robots interacting with humans in dynamic environments

RESEARCH OBJECTIVES

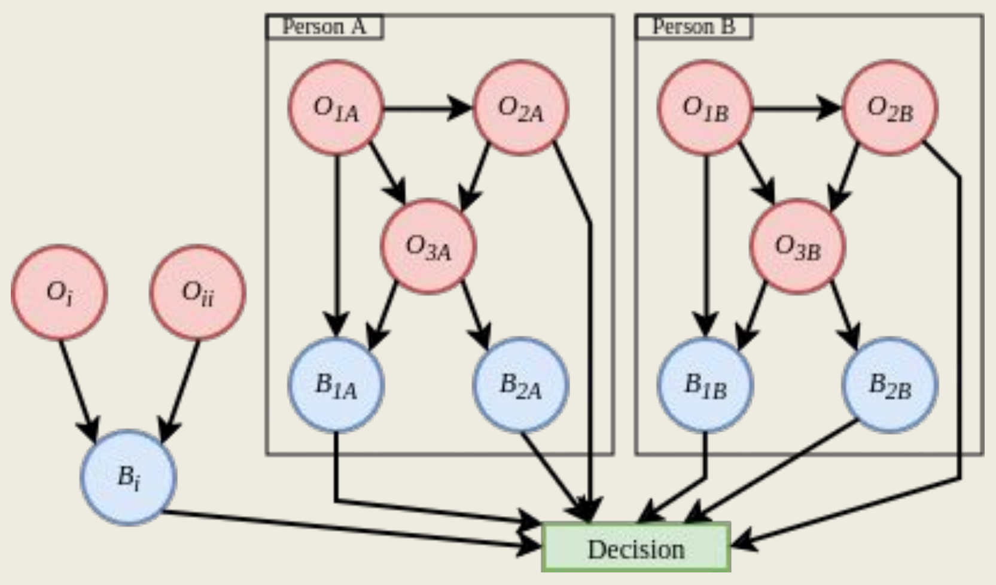
- **O1** - Generate explanations that can accurately **reflect** the **external environment** and **internal cognition**
- **O2** - Incorporate the effect of **uncertainty** on a robot's beliefs and decisions
- **O3** - Use **queries** from user to impose **conditions** on search for explanations
- **O4** - Investigate how best to **evaluate** the effect explanations have on the user's **understanding** of the robot's cognitive processes



Causal Model of external environment and internal cognition

Incorporate effect of **Uncertainty** in explanations

Impose conditions on search to match **User Queries**



"Why D_R = <A_R, T_R> and not D_H = <A_H, T_H>?"

Find intervention on S_R such that D_H is satisfied

Threshold for interventions to maximise discriminative power

I tried to talk to Person A because Person C was not very interested in me. If Person C was very interested in me, I would have tried to talk to them.

INITIAL RESULTS

[1] How do we generate **counterfactual** explanations for robot's in **dynamic**, **multi-person** environments using **causal models**?

We introduce a **two-layer perception** and **decision-making** system, model it with a causal model, and perform a **counterfactual search** with a discriminative condition based on **user queries**

[2] Do **explanations** of a robot's decision in a given context improve one's ability to **understand** (and thus **predict**) the decision-making of the robot in similar contexts?

We find that reasons obtained from counterfactual explanations **improve understanding**, but reasons + counterfactual statements **do not**, when compared to no explanations

Hypothesis: increased cognitive load due to longer explanations

FUTURE WORK

- Develop methods for incorporating uncertainty into explanations
  - Aleatoric vs Epistemic
  - Interventions on distributions rather than values
- Expand types of user queries
  - **Counterfactual**: "Why X and not Y?"
  - **Bi-factual**: "Why X at t_0 but Y at t_1?"
- More user studies:
  - How people understand uncertainty
  - Why counterfactual statements decreased prediction performance
- Develop and trial a query-explanation interface

Start date: October 2023
Research Plan defense: July 2024

TRAIL

Funding

HORIZON EUROPE MSCA
TRAIL (No 101072488)

Publications

[1] T. Love, A. Andriella and G. Alenyà. **Towards explainable proactive robot interactions for groups of people in unstructured environments**, 2024 ACM/IEEE International Conference on Human-Robot Interaction, 2024, Boulder, CO, USA, pp. 697–701.

[2] T. Love, A. Andriella and G. Alenyà. **What would I do if…? Promoting understanding in HRI through real-time explanations in the wild**, 33rd IEEE International Symposium on Robot and Human Interactive Communication, 2024, Pasadena, California, USA, IEEE, to appear.