Explainable Agency for Intelligent Autonomous Systems

Pat Langley, Ben Meadows, Mohan Sridharan

Computer Science / Electrical and Computer Engineering

University of Auckland, Auckland, NZ

Dongkyu Choi

Department of Aerospace Engineering

University of Kansas, Lawrence, KS USA

This research was supported by Grant N00014-15-1-2517. Thanks to Mike Barley, Ed Katz, and Peter Stone for useful discussions.

A Motivating Example

Consider a mission in which an autonomous robotic agent must:

- Deliver objects to a number of target offices;
- Collect any litter detected along the way;
- Avoid proximity of certain offices or people; and
- Retain enough battery charge to carry out these tasks.

Once the robot has completed its mission, it takes part in an 'after-action review'.

In this review, it must summarize its activities and answer questions about why it made various choices.

Previous Research

Intelligent systems that account for their decisions are not new:

- Early expert systems simply replayed reasoning chains, which led Swartout and Moore (1993) to call for better approaches.
- The most visible efforts (e.g., Ferrucci et al. 2010) have dealt with isolated decision tasks, not extended activities;
- Johnson (1994) and van Lent et al. (2004) reported agents for military missions that: recorded decisions made during missions, provided reasons on request, and handled counterfactual queries.
- However, both focused on knowledge-guided reactive execution rather than agent-generated plans.

Thus, there is some earlier work on which to build, but the field should devote far more attention to this problem.

Desirable Functions

Explainable agency will benefit from a number of functions:

- State alternatives considered during plan generation and reasons for making choices;
- Describe cases where execution diverged from the plan, how it responded, and reasons for taking these steps;
- Explain these reasons in terms of environmental states, mission objectives, and their relationships;
- Present reasoning about objectives in terms of both symbolic goals and numeric evaluation criteria;
- Present its activities at different levels of abstraction and detail, as appropriate to human queries.

We expect humans to exhibit these capacities, and they seem equally desirable in artificial agents.

Component Abilities

We can also specify abilities that enable these functions:

- Define *object categories* and *relations* in terms of observable percepts it can observe and link them to familiar words;
- Specify *mission objectives* as a set of symbolic goals with associated numeric utilities to communicate tradeoffs;
- Encode plans using *hierarchical structures* that decompose complex activities into increasingly finer subactivities;
- Record the *choices* made during planning, execution, and monitoring, including reasons for them, in episodic memory;
- Interpret *questions* about activities, use them to access relevant *memories*, and use retrieved content to *explain* activities.

We can implement these abilities in different ways, but they all seem needed for explainable agency.

Answering Questions

During the review, the agent must answer questions about its planning and execution like:

- What actions did you consider on coming to the intersection?
  - *Turning left or going straight ahead*.
- Which choice did you make and why did you make it?
  - *Turning left, since the path to John's office would be faster.*
- What did you expect to happen after you turned left?
  - *I would traverse a hallway with no obstacles.*
- What actually happened when you took that action?
  - *People came out of an office and stood in the hallway, so I backtracked and took the other route.*

Answering such questions means retrieving relevant content from episodic memory.

Explaining Use of Learned Expertise

Using learned expertise to control agents raises some special challenges for explainable agency, including:

- Communicating expected states during planning and inferred states during execution;
- Communicating reasons for selecting actions during planning and decisions about replanning during execution.

We can make use of learned expertise more explainable by:

- Learning mappings between internal representations and words in natural language.
- Using constrained notations (e.g., linear utility functions) that are easier to understand.

For more on this topic, see DARPA's XAI program (Gunning).

Summary Remarks

This talk identified a class of problems – *explainable agency* – that has received little attention from HRI researchers.

- Here an autonomous agent carries out extended missions, then answers questions about the reasons for its decisions.

I discussed some functions that arise in this task, important component abilities, and types of questions that can arise.

- The need for explainable agency is *not* primarily a learning issue, although it arises in that context as well.

In summary, we need more research in the AI community on this important and challenging topic.

End of Presentation