Explainable Agency in Human-Robot Interaction

Pat Langley
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA

1 Introduction and Motivation

The increasing use of robots in complex settings has led to a growing need for effective methods of interacting with such artificial agents. Most research in this area has focused on coordinating efforts between humans and robots during ongoing joint activities, but autonomous agents require different modes of engaging with humans that occur mainly before and after they carry out their missions.

We are interested in scenarios in which a human commander provides a robotic agent with a set of mission objectives, the robot enters the field to achieve those objectives, and finally returns after the mission is complete. The human then debriefs the robot by asking questions about its activities and its reasons for making decisions. These questions need not be in natural language, but communication must be in some format that is easy for humans to understand.

We will say that such robots are engaging in *explainable agency*. In this essay, we discuss some abilities that should be useful for computational artifacts that demonstrate this important capacity. The benefits of explaining one's past activities is not limited to robots; they hold for any agent that operates in some environment over time. This includes synthetic characters in virtual environments and systems that play games against others, but robotic agents are important case that we will use to illustrate the issues that arise.

2 Functions and Component Abilities

We claim that explainable agency depends relies on number of distinct but related functions that the autonomous systems can exhibit. We claim that an agent of this sort should:

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

We further claim that these explanatory functions benefit from a number of component abilities that let the agent:

- Define object categories and relations in terms of percepts it can observe and link them to familiar words, thus supporting descriptions of situations in encounters;
- Specify mission objectives as a set of symbolic goals with associated numeric utilities that let the agent communicate tradeoffs among alternatives;
- Encode plans using hierarchical structures that decompose complex activities into increasingly finer subactivities, enabling their at different levels of abstraction;
- Record the choices it makes on each step during plan generation, execution, and monitoring, including the reasons for selecting them, in an episodic memory;
- Interpret different types of questions about its activities, use them to access relevant portions of memory, and use the retrieved content to explain and justify its activities.

Taken together, these assumptions place high-level constraints on explainable agency. One can implement these tenets in different ways, but we maintain that any robot which exhibits the first set of abilities will also benefit from the second set of representations and processes.

These theoretical assumptions do not specify whether the expertise used during the agent's planning and execution is coded manually, learned from experience, or elements of both. However, they provide strong constraints on any formalism used for manual programming and on any learning mechanisms that acquire expertise automatically. In particular, they should combine qualitative and quantitative descriptions of states, refer to both symbolic goals and numeric utilities, and describe activities in hierarchical terms. This in turn suggests frameworks that we should consider when designing such explainable agents.

3 Representing Plans and Activities

Whether the expertise of our autonomous agent is handcrafted or acquired through learning, it must encode this expertise in forms that are well enough aligned with human cognition to enable effective communication. In this section,

we consider how our representational assumptions support the functions that we outlined earlier.

For example, although an autonomous robot has a clear need to describe its environment in precise quantitative terms for the purposes of recognition and control, humans refer to physical situations using words for abstract categories, such as *boulder* and *tree*, as well as abstract relations among them, such as *behind* and *between* (e.g., Burbridge & Dearden, 2012). Before it can report its activities and justify them, an explainable agent must define such qualitative concepts in terms of quantitative aspects of the environment encountered during its missions. These definitions may be encoded as logical formulae, probabilistic summaries, or even stored cases, but they should be associated with abstract symbols used to describe inferred and desired situations.

A similar issue arises in communicating the objectives themselves, which have both qualitative and quantitative aspects. Humans tend to specify mission objectives in terms of abstract goals, such as *place a movement sensor near the enemy camp* or *remain hidden from enemy radar*. As before, the explainable agent must ground these concepts in quantitative terms, but it must retain the ability to communicate them using abstract categories and relations. Moreover, such goals typically have different *utilities*, which themselves are encoded as numeric values or functions of the robot'senvironment. An explainable agent should not rely on a single utility for states it encounters; it should decompose this score in different elements, each associated with a symbolic goal (Langley et al., 2016). This factoring will also let it describe its activities in terms of tradeoffs among conflicting goals, say when it cannot remain fully hidden if it must come close enough to the enemy camp to deposit a sensor.

A final representational issue concerns people's tendency to describe activities in hierarchical terms, at different levels of abstraction. One response involves the robotic agent encoding both mission plans and execution traces in similar terms, with high levels denoting major steps and the lowest level describing primitive operations. Again, this will require symbolic structures of some form to encode the decomposition, similar to those in hierarchical task networks (Nau et al., 2003), along with terms that denote each activity or subactivity. However, even the highest levels of description may include quantitative features, such as distance traveled, fuel consumed, or average flying height. When associated with temporal aggregates, such numeric summaries can serve as important annotations to symbolic characterizations of the agent's activities.

4 Mission Planning and Execution

The autonomous robot in our scenario must carry out extended missions described by its human commander. The mechanisms that support this performance are constrained by the representational commitments that already discussed.

Planning will play an important role in most missions that involve extended activity. Building on our representational assumptions, an explainable agent should:

- Generate mental simulations of state sequences encoded in terms of both qualitative relations and quantitative attributes. This means not only predicting the numeric effects of actions, but also recognizing instances of categories and relations that arise in envisioned states. These are needed to describe situations considered in planning.
- Carry out search through the space of possible plans, selecting among alternative courses of action by reasoning about the symbolic goals that envisioned states satisfy, the numeric utilities associated with these goals, and which choices are better for the mission objectives.
- Decompose the planning task into subtasks that, when solved, produce a hierarchical plan that describes intended activity at different levels of temporal resolution. To this end, the agent can use a first-principles planner that explores a space of problem decompositions (Langley et al., 2016) or an HTN planner that uses domain-specific methods to constrain search (Nau et al., 2003).

The result of this process will be one or more plans that break the mission into subplans, that describe each anticipated state in both qualitative and quantitative terms, and that base choices on how alternative state sequences relate to goals and utilities.

Once the autonomous agent in our scenario has formulated a mission plan, it must then carry it out physically in the field. To this end, the system should:

- Execute the sequence of activities stored in the plan, determine the states these actions produce, and compare its expectations to these states. This means drawing on conceptual definitions to draw qualitative and quantitative inferences about its situation, which it can then use to monitor progress and detect anomalies.
- Determine whether anomalous events are sufficiently relevant and important to require replanning from the current situation. In many settings, revision will be necessary because the agent's original plan was based on incomplete or inaccurate information about the location and types of objects it encounters during execution.
- Upon determining that it must revise the current plan, deciding which elements to abandon, which to retain, and generating a new partial plan from which to initiate replanning. The agent should continue this process until it satisfies the mission's termination criterion.

Together, these mechanisms should let the robotic agent generate, execute, monitor, and revise plans for extended missions in which it must operate with little or no supervision. The plans it constructs will provide high-level guidance on its behavior, while monitoring and revision will let it adapt as needed to dynamic and unpredictable environments. Different settings will involve a different balance between these two activities, but seldom will either suffice alone.

5 Debriefing and Explanation

Once the robotic agent has completed a mission, the human commander should be able to debrief the system. This process might start with it providing a brief summary of events, but, more important, it can include answering questions about the reasons for the agent's decisions and actions.

Such capabilities will require the system to incorporate an episodic memory (Menager & Choi, 2016) that retains de-

tails about the mental and physical situations it has encountered. This should build on the representational and performance assumptions described earlier, so it should store:

- For each action or hierarchical method selected during plan construction, the goals it should achieve or maintain, the envisioned situation under which it would occur, the candidate actions that were considered, the scores associated with each alternative, and the final choice made;
- For each node in the search tree generated during planning, the times at which that partial plan was created and visited, decisions about its success or failure, and the reasons for each of these choices;
- For each step of a plan that the agent executed, the expected environmental situation, the observed state, and whether it deemed the descriptions were similar enough to continue executing the plan without revision; and
- For each point during execution at which the agent decided the observations had diverged from its expectations, which elements of the plan it retained, which ones it eliminated, and the reasons for these choices.

This content should be stored with the initial and revised plans themselves, not only because the latter provide natural scaffolds for such information, but also because the reasons for each decision were available to the agent at the time that it created these structures.

Of course, the system must also be able to access this information during after-mission debriefing. We assume the commander will have a trace of the robot's plans and activities, most likely organized hierarchically to reveal details only on demand, to serve as the basis for formulating questions. Whether he presents these in natural language, a graphical interface, or some other format, the agent should:

- Translate the question into a cue that specifies a point of interest during planning or execution in the same terms and syntax as structures in episodic memory;
- Use this cue to retrieve portions of episodic memory relevant to the query, asking for clarification when multiple results indicate ambiguity; and
- Present the retrieved answer to the human user in high-level terms, using a format appropriate to the question type and providing details only upon request.

The system should handle different forms of questions, from *What choices did you consider when you came to the brige?* to *What objectives did you hope to achieve by taking the left route?* to *Why did you select that alternative over other options?* These formats should map directly onto different types of content stored in episodic memory, with answers generated by filling in associated templates. The purpose is to give the commander insights into the agent's reasoning, not to provide him with entertaining responses.

Together, these capacities should let the robotic agent store the reasons for its decisions as it makes them, retrieve this information upon request, and provide justifications in terms that its commander will understand. Storing this content with mental structures generated during planning and execution of the mission offers natural ways to index and retrieve answers to queries stated in terms of its activities.

6 Closing Remarks

In this essay, we identified an important class of problems – *explainable agency* – that has received little attention with the human-robot interaction community. In these tasks, an autonomous agent carries out an extended mission in pursuit of human-specified objectives and, after completing it, must answer questions about the reasons for its decisions. We listed a number of functions that such a robotic agent should exhibit and a set of component abilities that we maintain will support these functions. The latter provide theoretical constraints on alternative approaches to explainable agency. These constraints are not *unique* to robotic systems, in that they apply equally well to any intelligent agent, whether physical or virtual, that communicates with humans. However, they are still *relevant* to human-robot interaction, and they merit attention from this community.

We discussed the three main elements – representation, plan generation and execution, and explanation during debriefing – in greater detail. However, our treatment was intentionally abstract, as the aim was not to describe a particular solution but rather a framework that solutions can utilize. We also took no position on whether the expertise used during planning, execution, and explanation should be programmed directly or learned from experience. Nevertheless, the theory suggests constraints on formalisms used to encode knowledge manually or to express the results of learning. Of course, the plan that we have proposed may require revision during efforts to carry it out, but it seems sufficient to let us take the first steps on this intriguing mission.

Acknowledgements

This work was supported by Grant N00014-15-1-2517 from the Office of Naval Research, which is not responsible for its contents. We thank Mike Barley, Dongkyu Choi, Ben Meadows, Stephanie Sage, Mohan Sridharan, and Peter Stone for helpful discussions on the ideas presented here.

References

Burbridge, C., & Dearden, R. (2012). Learning the geometric meaning of symbolic abstractions for manipulation planning. *Proceedings of Towards Autonomous Robotic Systems* (pp. 220–231). Springer.

Langley, P., Barley, M., Meadows, B., Choi, D., & Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Langley, P., Pearce, C., Bai, Y., Barley, M., & Worsfold, C. (2016). Variations on a theory of problem solving. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Menager, D. H. & Choi, D. (2016). A robust implementation of episodic memory for a cognitive architecture. *Proceedings of the Thirty-Eighth Annual Meeting of the Cognitive Science Society*. Philadelphia, PA.

Nau, D., Au, T., Hghami, O., Kuter, U., Murdock, J., Wu, D., & Yaman, F. (2003). SHOP2: An HTN planning system. *Journal of Artificial Intelligence Research*, *20*, 379–404.