Explainable Agency for Intelligent Autonomous Systems

Pat Langley* and **Ben Meadows*** and **Mohan Sridharan**‡
Department of Computer Science* / Electrical and Computer Engineering‡
University of Auckland, Private Bag 92019, Auckland 1142, NZ

Dongkyu Choi
Department of Aerospace Engineering
University of Kansas, Lawrence, KS 66045 USA

Explainable Agency

As intelligent agents become more autonomous, sophisticated, and prevalent, it becomes increasingly important that humans interact with them effectively. Machine learning is now used regularly to acquire expertise, but common techniques produce opaque content whose behavior is difficult to interpret. Before they will be trusted by humans, autonomous agents must be able to explain their decisions and the reasoning that produced their choices. We will refer to this general ability as *explainable agency*.

This capacity for explaining decisions is not an academic exercise. When a self-driving vehicle takes an unfamiliar turn, its passenger may desire to know its reasons. When a synthetic ally in a computer game blocks a player's path, he may want to understand its purpose. When an autonomous military robot has abandoned a high-priority goal to pursue another one, its commander may request justification. As robots, vehicles, and synthetic characters become more self-reliant, people will require that they explain their behaviors on demand. The more impressive these agents' abilities, the more essential that we be able to understand them.

Characterizing Explainable Agents

We will focus here on goal-directed autonomous agents engaged in tasks that require activity over time. We assume that such agents will generate plans using problem-space search, execute their plans in the environment, and adapt them as the need arises. As agents of this sort become more widespread, people will insist they be able to justify, or at least clarify, every aspect of these decision-making processes. We can specify the task of explainable agency more explicitly as:

- *Given* a complex set of objectives that require an agent's extended activity over time;
- *Given* background knowledge about categories, relations, and activities that are relevant to these objectives;
- *Produce* records of decisions made during plan generation, execution, and monitoring in pursuit of these aims;
- *Produce* summary reports, in human accessible terms, of the agent's mental and physical activities;
- *Produce* understandable answers to questions that are posed about specific choices and the reasons for them.

A relevant example involves an autonomous robot that plans and carries out a military mission, then participates in a debriefing session where it provides a summary report and answers directed questions from a human supervisor.

Explainable agency presents an important challenge for academia and industry. There has been substantial research on interactive synthetic characters and human-robot engagement, but it has emphasized joint activity during pursuit of ongoing tasks. We concentrate here on settings in which an agent receives instructions, carries them out with little or no interaction, and then describes and explains its decisions and actions afterwards. Providing such information after extended activity seems far more difficult than explaining individual choices as they arise, as in recommender systems.

Intelligent systems that account for their own decisions are not new. Some early expert systems simply replayed their reasoning chains, which led Swartout and Moore (1993) to call for more sophisticated explanation facilities. However, efforts in this area, including recent ones, have dealt mainly with individual decision-making tasks (e.g., Ferrucci et al. 2010) rather than the extended activities we expect autonomous agents to pursue. The most relevant work comes from Johnson (1994) and van Lent et al. (2004), who developed agents for tactical air combat and small unit settings, respectively, that recorded decisions made during missions, provided reasons on request, and dealt with counterfactual queries. However, both focused on knowledge-guided reactive execution rather than agent-generated plans.

We expand on Swartout and Moore's call to cover explainable agency, which we maintain requires four distinct functional abilities:

- The agent must be able to *explain decisions made during plan generation*. This should include stating the alternatives it considered, giving its reasons for selecting them over alternatives, and describing its expectations for each option. This can build on the foundations laid by recent work in explainable planning (e.g., Zhang, Zhuo, and Kambhampati 2015).
- The explainable agent must be able to *report which actions it executed*, presenting this information at different levels of abstraction as appropriate. The system should clarify how these actions relate to inferences it made, goals it adopted, and plans it generated. This is especially important in mine fields, the ocean, space, and other in-

hospitable environments where autonomous robots will be deployed and where transparency is essential.

- An autonomous agent must be able to *explain how actual events diverged from a plan and how it adapted in response*. It should also state on request the reasons for taking these steps, propose courses of action that seem better in hindsight, and even discuss what it would have done if other situations had arisen.
- An explainable agent must be able to *communicate its decisions and reasons* in ways that make contact with human concepts. This does not mean they must encode content in the same internal formalism, but the agent should present information in terms of beliefs, goals, and activities that people find familiar. These are often organized hierarchically, which should support both abstract accounts and drilling down on request.

Taken together, these four abilities provide the basics of explainable agency, but they say little about the quality of systems that have them. An autonomous agent that exhibits these features may still make a poor showing. This means we should do more than develop agents with such capabilities; we must also develop clear criteria for evaluating them.

Because explainable agency is motivated by our desire to understand the behavior of autonomous systems, human judgements should figure centrally in the evaluation process. These will probably include subjective ratings about suitability and clarity of answers to questions, as well as the degree of trust the agent engenders. More objective measures might include people's ability, after interaction, to predict an agent's behavior in future situations. Such metrics will let us distinguish between autonomous systems that explain themselves effectively and those that remain opaque.

Elements of Explainable Agency

The scientific and technical challenge we have posed is to create computational artifacts that behave well along the four functional abilities just outlined. We maintain that such explainable agents must incorporate three primary elements that can serve to guide research in this important area.

First, our target agents must *represent content that supports explanation*. This includes terms for domain concepts, relations, and activities that serve to describe states and actions in terms that people understand, as well as larger-scale structures for plans and execution traces. But it also includes notations for encoding choices that arise during search, selections made by the agent, and criteria used to make its decisions. Each of these levels seems likely to require both symbolic structures and numeric annotations.

Second, explainable agents must have an *episodic memory* that records states, actions, and values considered during plan generation, traces of plan execution in the environment, and anomalous events that led to plan revision. These should include the reasons for selecting some alternatives others, whether they focused locally on individual steps or globally on entire paths. These detailed memories are similar to the 'think-aloud protocols' that Newell and Simon (1972) collected from human problem solvers, but they are stored in memory for later access to support retrospective reports.

Finally, explainable agents must *access and extract content from episodic memory* to answer questions about their experiences. This requires an ability to interpret at least constrained natural language, use the result to identify and retrieve relevant structures from the episodic store, and report it in terms that humans find comprehensible. This process may be interactive, in that agents may ask for clarification to resolve ambiguity and humans may request elaboration about parts of answers. An agent's responses may use simple syntax, but they should describe choices it considered, selections it made, and reasons in understandable terms.

Addressing the Challenge

To date, there has been limited research on this important topic, despite the growing need for explainable agency in an increasingly AI-dependent world. We are not aware of any efforts that meet the four criteria presented earlier, which in turn motivated our proposal for three architectural components that, taken together, offer a path toward autonomous agents that justify their behavior. However, much of the relevant technology already exists: methods for automated planning, techniques for plan execution and monitoring, software for conversational question answering, and even approaches to storing and accessing episodic memories.

More generally, many existing AI research areas inform the crucial aspects of this challenge. This includes work on representing and organizing knowledge, using this knowledge for planning, execution, and communication, and acquiring this content through learning. The recent emphasis on statistical learning has made it more difficult to support explanation, so this issue deserves special attention. Nearly every subarea of AI has mature technology to offer, suggesting the time is ripe for work in this area, but we must still find ways to integrate it effectively and we must take seriously the need to communicate the reasons for agents' decisions to human partners. This will require substantial investment in research on explainable agency.

References

Ferrucci, D., et al. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine* 31: 59–79.

Johnson, W. L. 1994. Agents that learn to explain themselves. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1257–1263. Seattle, WA: AAAI Press.

Newell, A., and Simon, H. A. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Swartout, W. R., and Moore, J. D. 1993. Explanation in second generation expert systems. In J.-M. David, J.-P. Krivine, and R. Simmons, eds., *Second generation expert systems*. Berlin: Springer-Verlag.

Van Lent, M.; Fisher, W.; and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 900–907. San Jose, CA: AAAI Press.

Zhang, Y.; Zhuo, H. H.; and Kambhampati, S. 2015. Plan explainability and predictability for cobots. *https://arxiv.org/abs/1511.08158v1*.