Refinement-Based Architecture for Knowledge Representation, Explainable Reasoning and Interactive Learning in Robotics

Mohan Sridharan
School of Computer Science
University of Birmingham, UK
m.sridharan@bham.ac.uk

1 Motivation

Robots collaborating with humans in complex domains have to reason with different descriptions of incomplete domain knowledge and uncertainty. These descriptions include commonsense knowledge, e.g., default statements such as "textbooks are usually in the library" and "cereal boxes are typically in the kitchen", which hold true in all but a few exceptional circumstances. At the same time, information extracted by processing noisy inputs from sensors is often associated with quantitative measures of uncertainty, e.g., statements such as "I am 90% certain I saw the robotics book in the office". In addition, any robot operating in dynamic domains will have to augment or revise its existing knowledge over time. Furthermore, for effective collaboration with humans, robots should be able to explain their decisions, the underlying knowledge and beliefs, and the experiences that informed these beliefs. We have developed an architecture, REBA, which supports these capabilities by exploiting the complementary strengths of declarative programming, probabilistic planning, and interactive learning.

2 Architecture Overview

REBA, our refinement-based architecture for knowledge representation, explainable reasoning and interactive learning, is based on tightly-coupled transition diagrams at different resolutions. It may be viewed as a logician and statistician working together. Figure 1 shows an overview of the architecture. The different transition diagrams are described using an action language $\mathcal{AL}_d$ (Gelfond and Inclezan 2013), which has a sorted signature with statics, fluents and actions, and supports three types of statements, i.e., causal laws, state constraints, and executability conditions. We extend $\mathcal{AL}_d$ to support non-Boolean fluents and non-deterministic causal laws. We also expand the notion of the history of a dynamic domain to support prioritized defaults and define a model of such a history. Depending on the domain and tasks at hand, the robot chooses to plan and execute actions at two specific resolutions, but constructs explanations at other resolutions as needed. For ease of explanation, we will focus on two resolutions in our description here.

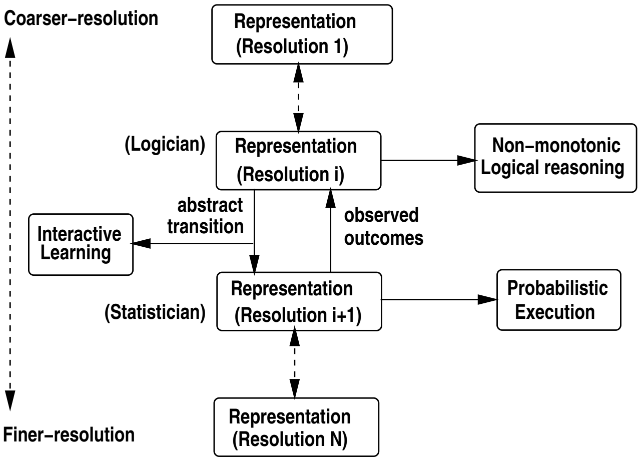
Coarser–resolution
Representation (Resolution 1)
(Logician)
Representation (Resolution i)
Non–monotonic Logical reasoning
abstract transition
observed outcomes
Interactive Learning
(Statistician)
Representation (Resolution i+1)
Probabilistic Execution
Finer–resolution
Representation (Resolution N)

Figure 1: Architecture supports representation and reasoning with tightly-coupled transition diagrams at different resolutions. It combines the strengths of declarative programming, probabilistic reasoning, and interactive learning.

Knowledge representation and reasoning: In the coarse-resolution, the robot represents and reasons with domain knowledge, including commonsense knowledge, at an abstract level. For example, a robot fetching objects in an office building would reason about places, objects, and default locations of objects. The fine-resolution transition diagram is then obtained by formally defining it as a *refinement* of the coarse-resolution transition diagram. This definition includes a *theory of observations* that models and reasons about the robot's ability to sense the values of domain fluents using knowledge-producing actions. In the context of fetching objects in an office building, the robot would (for example) now reason about grid cells in rooms and parts of objects, attributes that were previously abstracted away by the designer. Also, our definition of refinement ensures that for any given state transition in the coarse-resolution diagram, there is a path in the corresponding fine-resolution diagram between states that are refinements of the coarse-resolution states. In addition, the refined diagram is *randomized* to model non-determinism in action outcomes. For any given goal, the robot first performs non-monotonic logical reasoning at the coarse-resolution to compute a plan of *abstract actions*. In our architecture, this reasoning is achieved using *Answer Set Prolog*, a declarative programming paradigm (Gebser et al. 2012). Each abstract transition is then implemented as a sequence of concrete actions by automatically identifying and *zooming* to, and reasoning with, only the part of the fine-resolution transition diagram relevant to this coarse-resolution transition. Each concrete action is then executed by automatically generating relevant representations of probabilistic models of the uncertainty in perception and actuation. The outcomes of the action execution are added to the fine-resolution history, resulting in suitable entries being added to the coarse-resolution history. For more details about the representation and reasoning component, please see (Sridharan et al. 2019).

Interactive learning: Reasoning with incomplete domain knowledge can result in incorrect or suboptimal outcomes. It is possible to learn previously unknown actions and related axioms, but doing so in the most generic form may require many labeled examples in complex domains. It is difficult to provide such labeled examples in robot domains characterized by dynamic changes. Also, humans may not have the time and expertise to provide labeled examples or extensive feedback, and an action's effects may be immediate or delayed. Our architecture includes two schemes for interactive acquisition of labeled examples and knowledge: (i) active learning of actions and causal laws from human verbal descriptions of actions of other robots; and (ii) cumulative learning of action capabilities (i.e., affordances) and axioms using relational reinforcement learning and decision tree induction, based on observations from active exploration or reactive action execution. The key attribute of this learning approach is that reasoning with the existing knowledge informs and automatically limits interactive learning to the states, actions, and observations relevant to the task(s) and goal(s) at hand. For more details about the interactive learning component, please see (Sridharan and Meadows 2018).

Explainable reasoning: Our approach for explainable reasoning is based on a theory of explanations for human-robot collaboration. This theory comprises (i) claims about representing, reasoning with, and learning knowledge to support explanations; (ii) a characterization of explanations along three axes based on abstraction of representation, explanation specificity, and explanation verbosity; and (iii) a methodology for constructing explanations. This theory is implemented in our architecture by coupling the construction of explanations to the representation, reasoning and learning components summarized above. The robot receives explanatory questions in the form of verbal input from a human. This input is parsed using natural language processing tools and an underlying controlled vocabulary for human-robot interaction. The human user is then able to interactively obtain explanations at the desired level of abstraction, specificity, and verbosity. For more details about the theory of explanations and its implementation, please see (Sridharan and Meadows 2019).

Summary: Our architecture explores and exploits the interplay between knowledge representation, explainable reasoning, and learning, to address the corresponding challenges in human-robot collaboration. We have evaluated the capabilities of this architecture in simulation and on physical robots assisting humans in different tasks and domains. Experimental results indicate that our architecture supports reliable and scalable reasoning, learning, and explanations, in the presence of incomplete knowledge, violation of defaults, noisy observations, and unreliable actions.

Acknowledgments

This work was supported in part by the U.S. Office of Naval Research Science of Autonomy Award N00014-17-1-2434 and the Asian Office of Aerospace Research and Development award FA2386-16-1-4071. Opinions and conclusions reported in this paper are those of the author.

References

Gebser, M.; Kaminski, R.; Kaufmann, B.; and Schaub, T. 2012. *Answer Set Solving in Practice, Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan Claypool Publishers.

Gelfond, M., and Inclezan, D. 2013. Some Properties of System Descriptions of AL_d. *Journal of Applied Non-Classical Logics, Special Issue on Equilibrium Logic and Answer Set Programming* 23(1-2):105–120.

Sridharan, M., and Meadows, B. 2018. Knowledge Representation and Interactive Learning of Domain Knowledge for Human-Robot Collaboration. *Advances in Cognitive Systems* 7:77–96.

Sridharan, M., and Meadows, B. 2019. Theory of Explanations for Human-Robot Collaboration. In *AAAI Spring Symposium on Story-Enabled Intelligence*.

Sridharan, M.; Gelfond, M.; Zhang, S.; and Wyatt, J. 2019. REBA: A Refinement-Based Architecture for Knowledge Representation and Reasoning in Robotics. *Journal of Artificial Intelligence Research* 65:87–180.