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

Mohan Sridharan

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, often using data-driven methods. Furthermore, for effective collaboration with humans, robots should be able to describe their decisions, the underlying knowledge and beliefs, and the experiences that informed these beliefs. We have developed an architecture, REBA-KRL, which supports these capabilities by exploiting the complementary strengths and principles of step-wise refinement, non-monotonic logical reasoning, probabilistic planning, and interactive learning.

2 ARCHITECTURE OVERVIEW

Figure 1 is an overview of REBA-KRL, the refinement-based architecture for knowledge representation, explainable reasoning, and interactive learning, which is based on tightly-coupled transition diagrams at different resolutions. It may be viewed as a logician, statistician, and an explorer working together. The transition diagrams are described using an extension of action language \( \text{AL}_{\text{am}} \). The basic version has a sorted signature with statics, fluents, and actions, and supports three types of statements: causal laws, state constraints, and executability conditions, and the extension supports non-Boolean fluents and non-deterministic causal laws [5]. REBA-KRL also expands the notion of a history of a dynamic domain to support prioritized defaults [5]. Depending on the domain and tasks at hand, the robot chooses to reason, learn, and execute actions at two specific resolutions, but constructs on-demand descriptions of decisions, beliefs, and experiences in the form of relations between relevant objects, actions, and domain attributes at other resolutions. For ease of understanding, the description below focuses on two resolutions.

Knowledge representation and reasoning: The robot represents and reasons with incomplete commonsense domain knowledge, at

---

1 Intelligent Robotics Lab, School of Computer Science, University of Birmingham, UK, m.sridharan@bham.ac.uk
actions by automatically zooming to, and reasoning with, the part of the fine-resolution diagram relevant to the abstract transition. Each concrete action is then executed using algorithms that use learned probabilistic models of the uncertainty in perception and actuation. The outcomes of the fine-resolution action execution are added to the fine-resolution history, resulting in suitable entries being added to the coarse-resolution history and used for subsequent reasoning. Experimental results in simulation and on robots indicate reliable and efficient reasoning in complex domains [2][5].

Interactive learning: Reasoning with incomplete domain knowledge can result in incorrect or suboptimal outcomes. State of the art machine learning algorithms, especially deep learning algorithms, require a large number of labeled examples and considerable computational resources, which are often not available in many practical domains. REBA-KRL enables the robot to acquire knowledge of actions, action capabilities, and related axioms using three strategies: (i) human verbal descriptions of observed behavior; (ii) exploration of previously unexplored state transitions; and (iii) exploration of transitions that produce unexpected outcomes; these strategies are formulated as inductive learning or relational reinforcement learning (RLR) problems. *Reasoning and learning inform and guide each other*, enabling the automatic and efficient identification and use of only the relevant information to construct suitable mathematical formalisms (e.g., MDP for RLR) [6]. Figure 2 shows an example of the learning approach for the specific task of estimating the occlusion of objects and stability of object structures. State of the art methods use deep networks to extract image features and another deep network for making stability/occlusion decisions. In our case, features are extracted from any given input image and spatial relations between objects are grounded incrementally [3]. The agent first reasons with commonsense knowledge, image features, and spatial relations to make and relationally describe the occlusion and stability decisions. Relevant regions of interest are automatically extracted from images for which reasoning fails to make a decision, and used to train a deep network; these examples also induce constraints that are used for subsequent logical reasoning. Results indicate that this approach significantly improves the reliability and reduces computational effort in comparison with baseline deep network architectures—see [4][8].

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 as descriptions of decisions, beliefs, and experiences in the form of relations between relevant objects, actions, and domain attributes. This theory is implemented in REBA-KRL by coupling the construction of explanations to the representation, reasoning, and learning components summarized above. The robot receives requests or questions as (verbal) input from a human. This input is parsed using existing tools (e.g., for natural language processing) and an underlying controlled vocabulary for human-robot interaction. The human user is then able to interactively obtain relational descriptions at the desired level of abstraction, specificity, and verbosity. Experimental results indicate the applicability of this approach to different complex domains [7].

**Summary:** REBA-KRL exploits the interplay between knowledge representation, explainable reasoning, and interactive learning, to address key challenges in human-robot collaboration. These capabilities have been evaluated in simulation and on physical robots assisting humans in different tasks and domains, demonstrating reliable and scalable reasoning, learning, and explanation generation, in the presence of incomplete knowledge, violation of defaults, noisy observations, and unreliable actions.

**ACKNOWLEDGEMENTS**

REBA-KRL is the result of multiple research threads pursued in collaboration with Ben Meadows, Rocio Gomez, Tiago Mota, Heather Riley, Michael Gelfond, Jeremy Wyatt, and Shiqi Zhang. This work was supported in part by the U.S. Office of Naval Research Science and Systems Award N00014-13-1-0766 and N00014-17-1-2434, and the Asian Office of Aerospace Research and Development award FA2386-16-1-4071. All conclusions are those of the author alone.

**REFERENCES**


