24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

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

**Mohan Sridharan** [1]

## 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 $\mathcal{AL}_d$. 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 an abstract level, in the coarse-resolution. For example, a robot fetching objects in an office building would reason about places, objects, default locations of objects, and some axioms governing domain dynamics. This knowledge also includes a *adaptive theory of intentions* that incorporates principles of non-procrastination and persistence to respond to unexpected successes and failures. For example, if a robot plans to move two books, one at a time, from an office to the library and unexpectedly finds the second book in the library after moving the first one, it would stop executing the plan; if, after moving the second book to the library, it finds the first book missing, it will plan and execute actions to find and move the second book to the library. The domain's fine-resolution transition diagram is formally defined as a *refinement* of the coarse-resolution diagram. This definition includes a *theory of observations* that models the robot's ability to sense the values of domain fluents using knowledge-producing actions. Continuing with our example of fetching objects in an office building, the robot would now observe and reason about grid cells in rooms and parts of objects, attributes that were previously abstracted away by the designer. This definition of refinement guarantees 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 uncertainty in action outcomes. Then, for any given goal, the robot first computes a plan of *intentional abstract actions* using non-monotonic logical reasoning at the coarse-resolution. In REBA-KRL, this reasoning is achieved using *Answer Set Prolog*, a declarative programming paradigm [1]. Each abstract transition is implemented as a sequence of concrete

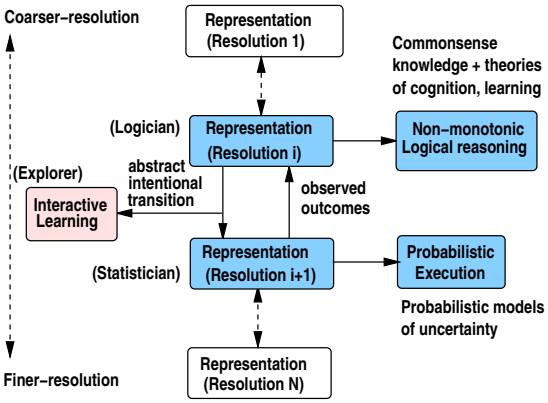


**Figure 1**: Architecture reasons with tightly-coupled transition diagrams at different resolutions, combining the strengths of declarative programming, probabilistic reasoning, and interactive learning.

[1] Intelligent Robotics Lab, School of Computer Science, University of Birmingham, UK, m.sridharan@bham.ac.uk

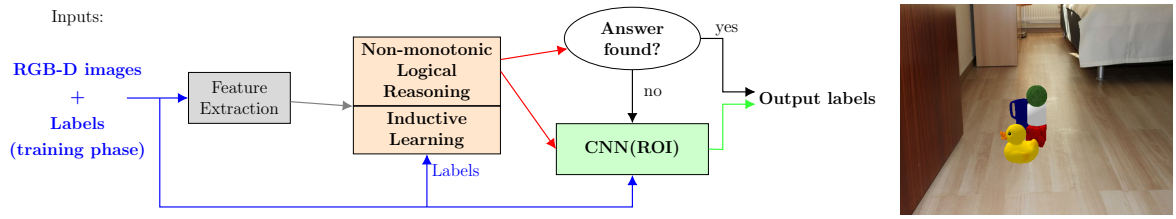


**Figure 2**: Architecture combining strengths of deep learning, inductive learning, and reasoning with commonsense knowledge, for scene understanding; example of simulated scene in which robot has to estimate occlusion of objects and stability of object structures.

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 (RRL) 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 RRL) [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 of Autonomy Awards 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

[1] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub, *Answer Set Solving in Practice, Synthesis Lectures on Artificial Intelligence and Machine Learning*, Morgan Claypool Publishers, 2012.
[2] Rocio Gomez, Mohan Sridharan, and Heather Riley, 'What do you really want to do? Towards a Theory of Intentions for Human-Robot Collaboration', *Annals of Mathematics and Artificial Intelligence, special issue on commonsense reasoning* (to appear), (2020).
[3] Tiago Mota and Mohan Sridharan, 'Incrementally Grounding Expressions for Spatial Relations between Objects', in *International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, (July 2018).
[4] Tiago Mota and Mohan Sridharan, 'Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning on Robots', in *Robotics Science and Systems*, Freiburg, Germany, (June 22-26, 2019).
[5] Mohan Sridharan, Michael Gelfond, Shiqi Zhang, and Jeremy Wyatt, 'REBA: A Refinement-Based Architecture for Knowledge Representation and Reasoning in Robotics', *Journal of Artificial Intelligence Research*, **65**, 87–180, (May 2019).
[6] Mohan Sridharan and Ben Meadows, 'Knowledge Representation and Interactive Learning of Domain Knowledge for Human-Robot Collaboration', *Advances in Cognitive Systems*, **7**, 77–96, (December 2018).
[7] Mohan Sridharan and Benjamin Meadows, 'Towards a Theory of Explanations for Human-Robot Collaboration', *Kunstliche Intelligenz*, **33**(4), 331–342, (December 2019).
[8] Mohan Sridharan and Heather Riley, 'Integrating Non-monotonic Logical Reasoning and Inductive Learning with Deep Learning for Explainable Visual Question Answering', *Frontiers in Robotics and AI, special issue on Combining Symbolic Reasoning and Data-Driven Learning for Decision-Making*, **6**, 125, (2019).