Method for Reasoning about other Agents' Beliefs from Observations

Alexander Nittka

Belief revision traditionally deals - from a first person perspective - with the question of what an agent should believe given an initial state and a revision input. This question is approached in two main ways: (i) formulating general properties a belief revision operator should satisfy and (ii) constructing specific revision operators. Reasoning about what another agent does in fact believe during a sequence of revisions is equally important. This third person perspective, which we look at in this thesis, has received much less attention so far.

We assume the observed agent to employ a particular framework for iterated non-prioritised revision. The task is to draw conclusions about the agent based on an observation containing information about which revision inputs the agent received and what it believed and did not believe upon receiving them. We are interested in conclusions concerning whether inputs are accepted or rejected and unrecorded beliefs. The general method will be to construct a potential initial epistemic state of the agent and progress the inputs recorded in the observation starting in that state. We call a state an explanation if it verifies the information contained in the observation. There are generally many explanations. In order to select one explanation, we will present and justify a set of preference criteria.

We introduce the assumed belief revision framework and show that any epistemic state defines a rational consequence relation. An observation can be translated into a partial description of such a relation. We can then make use of existing work on completing partial information about a rational consequence relation in order to construct an explanation. The explanation thus obtained is optimal with respect to the preference criteria.

In the first part of the work, we assume that the observation is complete in the sense that all revision inputs received during the time of observation are recorded and their logical content is completely known. These assumptions are essential to the optimality of the explanation. So far, this prevents us from dealing with cases where revision inputs are missed or where they are not all completely understood. The second part of the thesis investigates what can be said in such cases. We model unknown logical content by allowing the formulae recorded in the observation to contain unknown subformulae. Missing revision inputs can be dealt with by assuming the observation to contain additional entries where the revision inputs are formulae whose logical content is completely unknown. As we may not be informed about the positions or number of the additional inputs, further care needs to be taken when reasoning about the agent. We sketch algorithms for a number of cases differing in the detail of information available to the observer.

In the third part of the thesis, we look at the application of the methods in slightly different settings. Reasoning about different observations starting in the same state, which has applications in accessing expert knowledge or reasoning about software agents, is particularly interesting. We also consider variants of the assumed belief revision framework.