A Theory of Intentions for Intelligent Agents

Justin Blount
Southwest Research Institute
Michael Gelfond
Texas Tech University
Marcello Balduccini
Drexel University

September, 2015

The Paper Summary:

In the paper we

- Developed a theory allowing to reason about agents *intending to achieve a goal* and/or *intending to execute an activity* – a pair consisting of a goal and a plan aimed at its achievement.
- Utilized this theory for the development of methodology of building intelligent agents.

The work extends several previous papers on the subject authored by C. Baral and the authors.

It can be viewed as a precisely defined and executable refinement of classical BDI architecture based on recent achievements in knowledge representation.

Assumptions about Agents and their Environments

- We view the agent and its environment as a discrete dynamic system whose possible trajectories are represented by a state-action-state transition diagram.
- The agent is capable of making correct observations, remembering the domain history, and correctly recording the results of his attempts to perform actions.
- *Normally*, the agent is capable of observing the occurrence of all relevant exogenous actions.

Lifting some of these assumptions requires further investigation.

Challenges:

To achieve our goal we needed to decide how to:

- Represent the agent's knowledge about environment and its own capabilities and goals.
- Model the agent's beliefs.
- Define actions which may be intended by an agent given its beliefs and its goals.
- Define and implement an algorithm finding an intended action and prove its correctness.

Representing the Agent's Knowledge

Agent's knowledge is represented by a *domain description* $D_n = \langle T, \Gamma_n \rangle$ where

- T is a theory in action language $\mathcal{AL}$ describing all possible trajectories of the domain.
- Domain history Γ_n containing records of agent's actions and observations up to the current step n.

A *mental* part of a state of T consists of inertial fluents:

- $active\_goal(G)$ – *the agent intends to achieve* G.
- $status(m, k)$ – *the agent successfully executed first* k *elements of activity* m.

Theory of Intentions

The theory of intentions, included in action theories of intentional agents, can be viewed as a collection of axioms of $\mathcal{AL}$ defining the transformation of the agent's mental state.

It consist of 40 axioms of $\mathcal{AL}$, e.g.

$$start(M) \textbf{ causes } status(M, 0)$$

$$select(G) \textbf{ causes } active\_goal(G)$$

$$\neg active\_goal(G) \textbf{ if } main\_goal(G), G$$

Note, that mental state can be changed by mental actions like $start$ **and** $select$ **as well as by physical actions (e.g. an action making goal** G **true).**

Modeling the Agent’s Beliefs

Domain description $D_n = \langle T, \Gamma_n \rangle$ defines past trajectories of the system *believed to be possible by the agent*.

They are called *models* of D_n.

Note that the agent’s beliefs are *non-monotonic* – with growth of history new possible pasts may appear containing explanations of unexpected observations by unobserved exogenous actions.

However, mental fluents in all current states believed possible by the agent have the same values.

Intended Actions: The Definition

Agent's beliefs determine its intended actions as follows:

1. **If continuing execution of the ongoing activity may still lead to the goal then *execute the next action* of the activity.**
2. **If the goal is no longer active (either achieved or abandoned) then *stop* an ongoing activity.**
3. **If an ongoing activity is no longer expected to achieve its goal then *stop* the activity.**
4. **If the goal is active but no activity is selected to achieve it then *select and start* such an activity.**
5. **Otherwise *wait*.**

Intended Actions in Agent's Control Loop

Our, now *precisely defined*, notion of intended action is at the center of AIA control loop:

1. **Observe the world, interpret and record the result;**
2. **Find an intended action e;**
3. **Attempt to perform e and record the result;**
4. **Go to step 1.**

There are two major reasoning tasks:

- **Step (1) requires finding a model of agent's history. May need diagnostics.**
- **Step (2) requires finding an intended action. May need prediction (to check if the ongoing plan can succeed) and planning (if it does not or if the goal is new).**

Automating the Loop

Both reasoning tasks of an agent are reduced to computing answer sets of the CR-Program $\Pi(D_n)$ where $D_n = \langle T, \Gamma_n \rangle$ consisting of

- **Translation Π_T of action theory T into ASP.**
- **Π_M – rules for computing models of D_n.**
- **Π_I – rules for computing intended actions.**

1. **A model of D_n is found by computing an answer set A_1 of $\Pi(T, \Gamma_n)$.**
2. **An intended action e is extracted from an answer set A_2 of program**

   $$\Pi(T, \Gamma_n) \cup \{num\_of\_missed(x, n)\}$$

   where x is the number of unobserved exogenous actions; x is extracted from A_1.

Conclusions

We successfully tested the proposed methodology on a number of simple but non-trivial examples.

A demo of AIA architecture can be found on the TTU KRlab page.

The implementation of reasoning tasks in CR-Prolog allowed us to learn interesting things about power of this language.

The work should be expanded in many directions including allowing multiple intended goals, continuous or hybrid dynamic systems, weighted models, etc.

THANKS!