Temporal Modelling of Intentional Dynamics

Catholijn M. Jonker[1], Jan Treur[1] and Wouter C.A. Wijngaards[1]
[1]*Department of Artificial Intelligence, Vrije Universiteit Amsterdam,*
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
Email: <{jonker,treur,wouterw}@cs.vu.nl> URL: http://www.cs.vu.nl/{~jonker,~treur,~wouterw}

In this paper the internal dynamics of mental states based on beliefs, desires and intentions, is formalised using a temporal language. The use of a software environment to specify, simulate and analyse temporal dependencies between these intentional states in relation to behavioral traces is addressed.

1. INTRODUCTION

Dynamics has become an important focus within Cognitive Science in recent years; e.g., (Port & van Gelder, 1995). As one of the aspects, the dynamics of the interaction with the external world, and its implications for the representational content and dynamics of mental states have received attention; e.g., (Bickhard, 1993; Christensen & Hooker, 2000). Another important aspect is the internal dynamics of mental states, as can be found, for example in the dynamics of intentional notions (such as beliefs, desires and intentions) and their interaction with each other and with the external world. An example of a pattern for such internal dynamics is: if a desire and an additional reason (in the form of a belief about the world) to do some action are both present, then the intention to do the action is generated.

In this paper the internal dynamics of intentional mental states is addressed. It is shown how a temporal modelling environment can be used to specify, simulate and analyse models for these dynamics. A basic notion underlying the modelling is the notion of *functional role* or *profile*; e.g., in (Bickle, 1998) the functional profile of a mental state is considered as (pp. 205-206) '… a place in an abstract, systematically connected network running from sensory to behavior peripheries, in terms of the states and events that cause their occurrence and the subsequent states or events they cause.'

In this paper functional roles of intentional states are modelled in a *temporal language* in such a manner that causal relationships are formalised by temporal dependencies they entail. Since dynamics is a phenomenon occurring over real time, the real numbers are used as time frame. The temporal language can be used on the one hand for the *specification* of temporal relationships between intentional states and between intentional states and the external world. Such a temporal specification can be used to express a theory for these dynamics. On the other hand the language is the basis of a *software environment* that has been implemented and which can be used for the simulation and analysis of the internal intentional dynamics. In Section 2 the intentional notions on which the paper focuses are introduced. In Section 3 the formalisation for the dynamics is presented. An example and some results are presented in Section 4.

2. INTENTIONAL NOTIONS USED

The intentional notions from the BDI model (belief, desire and intention), are addressed in a static manner in e.g. (Rao & Georgeff, 1991; Linder, Hoek & Meyer, 1996); in our approach they are used in temporal perspective, see Figure 1. *Beliefs* are based on information that has been collected by observation of the external world in the present or in the past. Beliefs adapt to changes perceived in the external world. A belief denoted by the property belief(x, pos) means that the agent believes that property x holds.

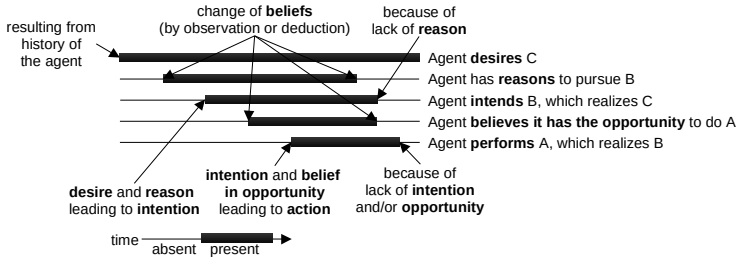

Figure 1. Intentional notions over time.

Desires are states of the world or changes to the world that are desired. From the set of desires that exist in a given situation some can be chosen to be pursued by creating an *intention* for them. An example of a pattern for such internal dynamics is: if a desire and an additional reason (in the form of a belief about the world) to do some action are both present, then the intention to do the action is generated. This intention lasts until the desire or the additional reason for it disappears. When the intention exists and it is believed that an *opportunity* presents itself, the *action* is performed in the external world.

3. DYNAMICAL FORMALISATION

In BDI-logics such as (Rao & Georgeff, 1991; Linder et al., 1996) internal processes are considered instantaneous. However, a more sincere formalisation is obtained if also internal processes extend over time. To be realistic, time has to be real, not measured in computational steps. In our formalisation, real-time temporal relationships are defined that take into account the delay between cause and effect, together with the durations of those cause and effect situations. In the following the term *agent* is used to refer to the subject and *system* is used to refer to the agent and the external world together. Intervals of real numbers are denoted like: $[x, y)$ meaning $\{p \in \mathbb{R} \mid p \geq x \wedge p < y\}$. Thus, '[' or ']' stands for a closed end of the interval, and '(' or ')' stands for an open end of the interval.

3.1. State Properties

The states of the system are characterised by *state properties*. State properties are formalised using (logical) formulae over a specific ontology. For an ontology Ont, the set of *atoms* AT(Ont) contains the atomic properties expressed in terms of the ontology. The set of *state properties* SPROP(Ont) contains all the propositional formulas built out of the atoms using standard propositional connectives. More specifically, the following ontologies are used. Firstly, *world state properties*

express properties of a particular situation in the material world, using ontology EWOnt. Secondly, the *internal physical state properties* of the agent are expressed using IntOntP. The combined physical ontology is OntP $=_{def}$ EWOnt $\cup$ IntOntP. Thirdly, the ontology for internal mental state properties is denoted by IntOntM. The ontology for all state properties is denoted by AllOnt $=_{def}$ EWOnt $\cup$ IntOntP $\cup$ IntOntM.

3.2. States

a) A *physical state* P of the system is an assignment of truth values {true, false} to the set of physical state atoms AT(OntP) of the system. The set of all possible physical states is denoted PS.

b) A (partial) *mental state* M of the system is an assignment of truth values {true, false, unknown} to the set of internal mental state atoms, AT(IntOntM). The set of all possible mental states is denoted by MS. Three valued states are used to avoid commitment to closed world assumptions or explicit specification of negative conclusions.

c) At each time-point the system is in one state. This state is from the set States $=_{def}$ PS x MS.

d) The *satisfaction relation* $s \models \varphi$ between states and state properties means that property φ holds in state s.

3.3. Traces

The system when viewed over a period of time, will produce several states consecutively. A function $\mathcal{T}$ returning the state for each time point is called a *trace*, $\mathcal{T}$: $\mathbb{R} \rightarrow$ States. The notation state($\mathcal{T}$, t, m), where $\mathcal{T}$ is a trace, $t \in \mathbb{R}$ and $m \in$ {physical, mental}, means the physical or mental state at time t in trace $\mathcal{T}$. The notation state($\mathcal{T}$, t) is by definition $\mathcal{T}(t)$. The set of all possibly occurring traces is denoted $\mathcal{W}$.

The behaviour of the agent and its environment is defined by a set of traces. To specify such a set of traces, temporal relationships between the state properties over time are defined, which express certain constraints on the relative timing of the occurrence of state properties within a trace.

3.4. The '$\rightarrow\!\!\!\rightarrow$' Relation and the '$\bullet\!\!-$' Relation

Let $\alpha, \beta \in$ SPROP(AllOnt). The state property α *follows* state property β, denoted by $\alpha \rightarrow\!\!\!\rightarrow_{e, f, g, h} \beta$, with *time delay interval* [e, f] and *duration parameters* g and h if

$\forall \mathcal{T} \in \mathcal{W}\ \forall t1:\quad [\forall t \in [t1 - g, t1) : \text{state}(\mathcal{T}, t) \models \alpha \ \Rightarrow\ \exists d \in [e, f]\ \forall t \in [t1 + d, t1 + d + h)\ : \text{state}(\mathcal{T}, t) \models \beta\ ]$

Conversely, the state property β *originates from* state property α, denoted by $\alpha \bullet\!\!-_{e, f, g, h} \beta$, with time delay in [e, f] and duration parameters g and h if

$\forall\ \mathcal{T} \in \mathcal{W}\ \forall$ t2:$\quad [\forall t \in [t2, t2 + h)\ : \text{state}(\mathcal{T}, t) \models \beta\ \Rightarrow\ \exists d \in [e, f]\ \forall t \in [t2 - d - g, t2 - d)\ \text{state}(\mathcal{T}, t) \models \alpha]$

If both $\alpha \rightarrow\!\!\!\rightarrow_{e,f,g,h} \beta$, and $\alpha \bullet\!\!-_{e,f,g,h} \beta$ hold, this is denoted by: $\alpha \bullet\!\!\rightarrow\!\!\!\rightarrow_{e,f,g,h} \beta$. It means that α *leads to* β; β will occur after α has happened. The relationships between the variables α, β, e, f, g, h, t0, t1 and t2 are depicted in Figure 2.

Notice that when φ holds for a continued length of time and $\varphi \bullet\!\!\rightarrow\!\!\!\rightarrow_{e,f,g,h} \psi$ and $e + h \geq f$, that then ψ will also hold for a continued length of time, without gaps.

4. AN EXAMPLE FORMALISATION

In order to demonstrate the formalisation and automated support presented in this paper, a simple example description is presented. In this example, the test subject is a common laboratory mouse, that is presented with cheese. Mostly, the mouse will try to eat the cheese, but a transparent screen can block access to the cheese. An example formalisation is:

..... Sensing .....

hungry $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(hungry, pos) $\wedge$ $\neg\beta$(hungry, neg).

$\neg$hungry $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(hungry, neg) $\wedge$ $\neg\beta$(hungry, pos).

cheese_present $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(cheese_present, pos) $\wedge$ $\neg\beta$(cheese_present, neg).

$\neg$cheese_present $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(cheese_present, neg) $\wedge$ $\neg\beta$(cheese_present, pos).

screen_present $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(screen_present, pos) $\wedge$ $\neg\beta$(screen_present, neg).

$\neg$screen_present $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ β(screen_present, neg) $\wedge$ $\neg\beta$(screen_present, pos).

..... Internal Processes .....

β(hungry, pos) $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ δ(eat_food).

δ(eat_food) $\wedge$ ρ_1 $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ ι(eat_cheese).

ι(eat_cheese) $\wedge$ o_1 $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ α(eat_cheese).

ρ_1 = β(cheese_present, pos).

o_1 = β(screen_present, neg).

..... World Processes .....

α(eat_cheese) $\wedge$ cheese_present $\bullet\!\!\rightarrow\!\!\!\!\rightarrow_{1,\,5,\,10,\,10}$ $\neg$hungry.

The graph in Figure 3 shows the reaction of the mouse to changes in the environment. Time is on the horizontal axis. The world state properties and the intentional notions are listed on the vertical axis. The parameter λ is fixed at 0.25. A dark box on top of the line indicates the notion is true, and a lighter box below the line indicates that the notion is false.

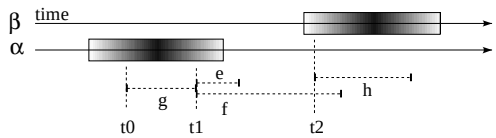

Figure 2. The time relationships between variables.

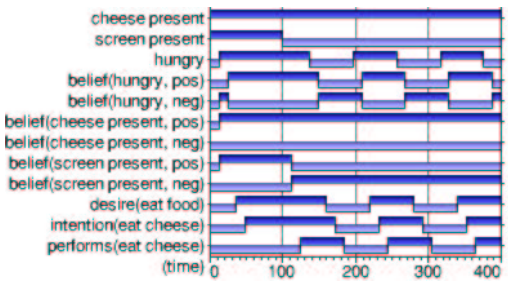

Figure 3. Results when the environment is set initially to have cheese and a screen. Later the screen is removed.

As can be seen, the mouse is not hungry at the very start, but quickly becomes hungry. It desires to eat the cheese, and intends to do so, but the screen blocks the opportunity to do so. When the screen is removed, the mouse eats. After a while it stops eating, as it is not hungry anymore. Subsequently it enters a cycle where it becomes hungry, eats, and becomes hungry again.

5. DISCUSSION

This paper addresses formalisation of the internal dynamics of intentional states, i.e. states involving beliefs, desires and intentions. In available literature on formalisation of intentional behaviour, such as (Rao & Georgeff, 1991; Linder et al., 1996) the internal dynamics of intentional mental states are ignored. The formalisation of the internal dynamics of intentional states introduced in this paper is based on a real time temporal language. Within this (quite expressive) temporal language a specific format is defined which can be used to specify temporal relationships that describe (constraints on) the dynamics of intentional states and their interaction with the external world. Specifications in this specific format have the advantage that they can be used to perform simulation, based on the paradigm of executable temporal logic (Barringer et al., 1996). The approach subsumes discrete simulation, for example as performed in Dynamical Systems Theory (Port & van Gelder, 1995) as a special case (with $e=f=1$ and $g=h=0$).

A software environment has been implemented including three programs. The first simulates the consequences of a set of temporal relationships of intentional states over time. The second program interprets a given trace of intentional states over time (in terms of beliefs, desires and intentions), and makes an analysis whether the temporal relationships hold, and, if not, points at the discrepancies. A third program takes into account physical states and their (possible) relation to intentional mental states. Physical traces, for example obtained by advanced scanning techniques, can be input and analysed with respect to possible interpretations in terms of intentional mental states.

REFERENCES

Barringer, H., M. Fisher, D. Gabbay, R. Owens, & M. Reynolds (1996). *The Imperative Future: Principles of Executable Temporal Logic*, Research Studies Press Ltd. and John Wiley & Sons.

Bickhard, M.H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, pp. 285-333.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. MIT Press, Cambridge, Massachusetts.

Christensen, W.D. & C.A. Hooker (2000). *Representation and the Meaning of Life*. In: (Clapin et al., 2000).

Clapin, H., Staines, P. & Slezak, P. (2000). *Proc. of the Int. Conference on Representation in Mind: New Theories of Mental Representation*, 27-29th June 2000, University of Sydney. To be published by Elsevier.

Dretske, F.I. (1991). *Explaining Behaviour: Reasons in a World of Causes*. MIT Press, Cambridge, Massachusetts.

Linder, B. van, Hoek, W. van der & Meyer, J.-J. Ch. (1996). How to motivate your agents: on making promises that you can keep. In: Wooldridge, M.J., Müller, J. & Tambe, M. (eds.), *Intelligent Agents II. Proc. ATAL'95* (pp. 17-32). Lecture Notes in AI, vol. 1037, Springer Verlag.

Port, R.F. & Gelder, T. van (eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, Massachusetts.

Rao, A.S. & Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-architecture. In: (Allen, J., Fikes, R. & Sandewall, E. ed.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (KR'91), Morgan Kaufmann, pp. 473-484.