Temporal and Interactionist Perspectives on the Dynamics of Mental States

Catholijn M. Jonker and Jan Treur
Department of Artificial Intelligence, Vrije Universiteit Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
Email:{jonker, treur}@cs.vu.nl URL: http://www.cs.vu.nl/{~jonker,~treur}

This paper contributes trace semantics and a temporal trace language to provide a grounding and formalisation of the dynamics of mental states in relation to the dynamics of the interaction with the external world. The approach provides a foundation for the dynamical and interactionist perspective on cognitive phenomena.

1. INTRODUCTION

In recent literature on cognitive science and philosophy of mind, perspectives on cognitive functioning are proposed, where dynamics and interaction with the environment are central; e.g. (Bickhard, 1993; Clark, 1997; Port and van Gelder, 1995; Clapin et al., 2000; Christensen and Hooker, 2000). For example, (Bickhard, 1993) emphasises the relation between the (mental) state of a system (or agent) and its past and future in the interaction with its environment:

> 'When interaction is completed, the system will end in some one of its internal states - some of its possible final states. (..) The final state that the systems ends up in, then, serves to implicitly categorise together that class of environments that would yield that final state if interacted with. (..) The overall system, with its possible final states, therefore, functions as a *differentiator* of environments, with the final states implicitly defining the differentiation categories. (..) Representational content is constituted as indications of potential further interactions. (..) The claim is that such differentiated functional indications in the context of a goal-directed system constitute representation - emergent representation.'.

This suggests that mental states need to be grounded in interaction histories on the one hand, and have to be related to future interactions on the other hand. Some of the questions addressed in this paper are the following. What exactly is an interaction history? How does this precisely relate to a mental state? What about future traces, if they depend as well on the environment' s dynamics? How do they relate to mental states? To answer these questions, the temporal aspect of the dynamics of mental states and the interaction with the environment is formalised in this paper on the basis of formally defined traces and an expressive temporal trace language.

First, as a basis, in Section 2 the notion of trace and the temporal trace language are defined. In Section 3 it is shown how internal state properties can be formally related to sets of interaction traces to obtain their representational content or semantics. Section 4 addresses how these sets of traces can be characterised by formulae in the temporal trace language. Criteria are identified that express when a temporal formula defines a class of interaction traces that can be related to a specific mental property. Such a temporal formula can be viewed as a temporal grounding or temporal representation of this mental property. Section 5 concludes the paper with a brief discussion.

2. DYNAMICS, TRACES AND TRACE LANGUAGE

In this section dynamics is formalised by states that change over time. A *state* for a state ontology Ont (i.e., names for state properties) is an assignment of truth-values from {true, false} to the set of ground atoms At(Ont). The *set of all possible states* for ontology Ont is denoted by STATES(Ont). In particular, STATES(OvOnt) denotes the set of all possible *overall states* (of the agent, its body and external world together); STATES(IntOnt) is the set of all of the agent's possible *internal states*. Moreover, STATES(InOnt) is the set of possible input states (e.g., sensor states), and STATES(OutOnt) the set of output states (e.g., effector states); STATES(InterfaceOnt) denotes the set of all *interface states* (i.e., input and output together) between agent and external world. The standard satisfaction relation between states and state properties is used: $S \models p$. means that property p holds in state S. For a state S over ontology Ont with sub-ontology Ont', a restriction of S to Ont' can be made, denoted by $S|Ont'$; this restriction is the member of STATES(Ont') defined by $S|Ont'(a) = S(a)$ if $a \in At(Ont')$. For example, if S is an overall state, i.e., a member of STATES(OvOnt), then the restriction of S to the internal atoms, $S|IntOnt$ is an internal state, i.e., a member of STATES(IntOnt).

To describe behaviour of the agent, explicit reference is made to time in a formal manner. A fixed *time frame* T is assumed which is linearly ordered. Depending on the application, it may be dense (e.g., the real numbers), or discrete (e.g., the set of integers or natural numbers or a finite initial segment of the natural numbers), or any other form, as long as it has a linear ordering.

A *trace* $\mathcal{M}$ over an ontology Ont and time frame T is a mapping $\mathcal{M}: T \rightarrow STATES(Ont)$, i.e., a sequence of states $\mathcal{M}_t$ ($t \in T$) in STATES(Ont). The set of all traces over ontology Ont is denoted by TRACES(Ont), i.e., $TRACES(Ont) = STATES(Ont)^T$. A *temporal world description* $\mathcal{W}$ is a set of traces over the overall ontology, i.e., $\mathcal{W} \subseteq TRACES(OvOnt)$. This set represents all possible developments over time (respecting the world's laws) of the agent and part of the world considered. This set formalises the processes of the agent, its body and the world as one integrated whole, as advocated in (Clark, 1997): 'Putting Brain, Body and World Together Again'. Given a trace $\mathcal{M}$ over the overall ontology OvOnt, the state of the input interface at time point t, i.e., $\mathcal{M}_t|InOnt$, is denoted by $state(\mathcal{M}, t, InOnt)$. Analogously, $state(\mathcal{M}, t, OutOnt)$ denotes the state of the output interface of the agent, and $state(\mathcal{M}, t, IntOnt)$ the internal state at time point t.

To focus on different aspects of the agent and time, traces can be restricted to a specific ontology Ont and time interval Interval. The ontology parameter Ont indicates which parts of the agent or world are considered. For example, when this parameter is InOnt, then only input information is present in the restriction. The time interval parameter Interval specifies the part of the time frame of interest. The *restriction* $\mathcal{M}_{Interval}^{Ont}$ of a trace $\mathcal{M}$ to time in Interval and information based on Ont is a mapping $\mathcal{M}_{Interval}^{Ont}: Interval \rightarrow STATES(Ont)$, defined by: $\mathcal{M}_{Interval}^{Ont}(t) = \mathcal{M}(t)|Ont$ if $t \in Interval$. For example, the *interaction trace* $\mathcal{M}_{\leq t}^{InterfaceOnt}$ denotes the restriction of $\mathcal{M}$ to the past up to t and to interface atoms.

Comparable to the approach in situation calculus, the sorted predicate logic temporal trace language TTL is built on atoms referring to, e.g., traces, time and

state properties, such as state($\mathfrak{M}$, t, OutOnt) |= p. Here |= is a predicate symbol in the language, comparable to the Holds-predicate in situation calculus. Temporal formulae are built using the usual logical connectives and quantification (for example, over traces, time and state properties). The set TFOR(Ont) is the set of all *temporal formulae* that only make use of state ontology Ont. A *past formula* for $\mathfrak{M}$ and t is a temporal formula $\psi(\mathfrak{M}, t)$ such that each time variable different from t is restricted to the time interval before t. In other words, for every time quantifier for a variable s a restriction of the form $s \leq t$, or $s < t$ is required within the formula. The set of past formulae over ontology Ont w.r.t. time point t is denoted by PFOR(Ont, t). Note that for any past formula $\psi(\mathfrak{M}, t)$ it holds:

$$\forall \mathfrak{M}, \mathfrak{N} \in \mathfrak{W} \forall t \, [\, \mathfrak{M}_{\leq t} = \mathfrak{N}_{\leq t} \Rightarrow [\, \psi(\mathfrak{M}, t) \Leftrightarrow \psi(\mathfrak{N}, t) \,] \,].$$

Similarly, FFOR(Ont, t) denotes the set of future formulae over ontology Ont w.r.t. time point t: every time quantifier for a variable s is restricted by $s \geq t$ or $s > t$.

3. INTERNAL STATE PROPERTIES AND INTERACTION DYNAMICS

As put forward in the introduction, according to the interactionist view, a possible internal state '… serves to implicitly categorise together that class of environments that would yield that final state if interacted with', cf. (Bickhart, 1993). Using our framework introduced in Section 2 the set of interaction histories over ontology Ont leading to internal property p, is formally defined by

$$\text{PTRACES}(\text{Ont}, p) = \{ \mathfrak{M}_{\leq t}^{\text{Ont}} \mid t \in T, \mathfrak{M} \in \mathfrak{W}, \text{state}(\mathfrak{M}, t, \text{IntOnt}) \models p \}$$

Besides, the way in which internal properties themselves lead to particular possible types of future interactions is also crucial for their meaning (Bickhard, 1993). Therefore, for an internal state property p, and an ontology Ont, the set of *all future traces* for t over Ont *allowed by* p is defined by:

$$\text{FTRACES}(\text{Ont}, p) = \{ \mathfrak{M}_{\geq t}^{\text{Ont}} \mid t \in T, \mathfrak{M} \in \mathfrak{W}, \text{state}(\mathfrak{M}, t, \text{IntOnt}) \models p \}$$

Based on these formal definitions, the *representational content of an internal state property* p is defined as the pair of sets PTRACES(InterfaceOnt, p), FTRACES(InterfaceOnt, p).

The concepts introduced are illustrated with a example of the internal state property pain. This property is assumed to have relationships to the input properties injury and heat. Each of injury or heat causes pain (1), and they are the only possible causes (2). The set of world traces $\mathfrak{W}$ for this example reflects this in the sense that for any trace, always after injury occurs at the input, the internal state property pain will occur further on in the trace, and the same holds for heat at the input (1). Moreover, if pain occurs in a trace, then earlier in the trace one of or both heat and injury occurred at the input (2). An example of an interaction history on the *input* of the agent leading to pain, i.e., an element of PTRACES(InOnt, pain), is the following (partially depicted) interaction trace:

t0. injury: false,heat: false ; t1. injury: true,heat: false ; t2. injury: true,heat: false

Note that in such a trace a delay may occur between the occurrence of the sensory input and the occurrence of the internal state property pain. How much delay ≥ 0 is taken into account is easily expressible in the temporal approach introduced by taking the real numbers as time frame.

For the future perspective, the internal state property pain is assumed to have relationships to the output properties move and ouch!. The property pain

unconditionally causes the cry ouch!, and causes the action move depending on whether or not the environment object that caused the pain stays close or returns (e.g., a wasp) (1). It is assumed that these outputs are only generated if the internal state property pain holds (2). The set of world traces σ_W reflects this in the sense that always after a time point where pain occurs in the internal state, (a) the output property ouch! will occur further on in the trace, and (b) if later on at the input present occurs, i.e., the object causing the pain is still there or returned, then this is followed by move at the output later on in the trace (within a certain time d, which for simplicity will be left out). Note that this implies learned behaviour: e.g., all wasps encountered in future will trigger an avoidance reaction. An example of a (partially depicted) interaction future allowed by pain, i.e., in FTRACES(InterfaceOnt, pain), is as follows:

t0. present: false ; t1. present: true ; t2. move: true

4. TEMPORAL REPRESENTATIONS OF MENTAL STATES

Until now the interaction histories and futures have been formally defined by semantic set-theoretic means. However, using the temporal trace language introduced in Section 2, sets of traces can be characterised by temporal formulae as well. In the pain example the set PTRACES(InOnt, pain) is characterised by:

$\mathcal{M}_{\leq t}^{InOnt} \in$ PTRACES(InOnt, pain) $\Leftrightarrow$ $\psi_{P(\mathcal{M}, t)}$

where $\psi_{P(\mathcal{M}, t)} \in$ PFOR(InterfaceOnt, t) is the past formula

$\exists t1 \leq t$ state($\mathcal{M}$, t1, InOnt) |= injury $\vee$ $\exists t2 \leq t$ state($\mathcal{M}$, t2, InOnt) |= heat.

The formula $\psi_{P(\mathcal{M}, t)}$ can be considered as an *external temporal representation* of the internal state property pain. If a nonzero delay of at least d is taken into account, $\psi_{P(\mathcal{M}, t)}$ has to be replaced by $\psi_{P(\mathcal{M}, t-d)}$, to make the equivalence hold. If a delay with some randomness between 0 and d is assumed, then $\psi_{P(\mathcal{M}, t)}$ can be replaced by $\exists d' \, 0 \leq d' \leq d \, \& \, \psi_{P(\mathcal{M}, t-d')}$, to guarantee the implication $\Rightarrow$. However, the implication $\Leftarrow$ then does not hold.

For the future, to characterise the set of traces FTRACES(InterfaceOnt, pain) in the form

$\mathcal{M}_{\geq t}^{InterfaceOnt} \in$ FTRACES(InterfaceOnt, pain) $\Leftrightarrow$ $\psi_{F(\mathcal{M}, t)}$

a candidate formula $\psi_{F(\mathcal{M}, t)} \in$ FFOR(InterfaceOnt, t) is:

$\exists t1 \geq t$ state($\mathcal{M}$, t1, OutOnt) |= ouch! & $\forall t2 \geq t$ [state($\mathcal{M}$, t2, InOnt) |= present $\Rightarrow$ $\exists t3 \geq t2$ state($\mathcal{M}$, t3, OutOnt) |= move]

Also here a zero or fixed delay is assumed. This guarantees the implication $\Rightarrow$. However, an additional problem here is caused by the conditional in $\psi_{F(\mathcal{M}, t)}$. Traces may occur where never the condition on present comes to hold. Then the implication is trivially true. But it is not satisfactory on this basis to conclude by the implication $\Leftarrow$ that there has been pain (although in the example, also the ouch! property may play a role, but this does not solve the principle of the problem).

The pain example illustrates that temporal formulae characterising the representational content of an internal notion (in the sense of the past and future traces sets), in a simple manner depends on two assumptions: (1) fixed delay (2) no conditionals. Moreover, a silent assumption was (3): within the state ontology an internal state property exists for the considered notion (pain). For a mathematical modelling approach, assumption (1) is customary (although not

quite desirable), so one could live with that. From an interactionist perspective, assumption (2), however, is unacceptable, because it excludes the possibility of the agent to let its behaviour in the future interactively depend on conditions that may or may not occur in a specific future interaction trace. Assumption (3) is innocent in the study of internal state properties and their content. However, if the attribution of mental properties based on observed behaviour is addressed, then assumption (3) would be artificial.

To avoid the assumptions discussed, the following notions are introduced. Let $\psi_{P(\mathcal{M}, t)} \in$ PFOR(InterfaceOnt, t) be a past formula and $\psi_{F(\mathcal{M}, t)} \in$ FFOR(InterfaceOnt, t) a future formula over the interface ontology. Moreover, let Ont be a given ontology (e.g., the internal ontology), and $\varphi_{(\mathcal{M}, t)} \in$ TFOR(Ont) a temporal formula over Ont.

The past formula $\psi_{P(\mathcal{M}, t)}$ is a *sufficient past interaction grounding for* $\varphi_{(\mathcal{M}, t)}$ if*:*

$\forall \mathcal{M} \in \mathfrak{M} \forall t [\ \forall \mathcal{N} \in \mathfrak{M} [\ \mathcal{M}_{\geq t} = \mathcal{N}_{\geq t} \Rightarrow \exists t1 \leq t\ \ \psi_{P(\mathcal{N}, t1)}\ ]\ \Rightarrow\ \exists t2 \geq t\ \varphi_{(\mathcal{M}, t2)}\ ]$

The past formula $\psi_{P(\mathcal{M}, t)}$ is a *necessary past interaction grounding for* $\varphi_{(\mathcal{M}, t)}$ if*:*

$\forall \mathcal{M} \in \mathfrak{M}\ \forall t [\ \varphi_{(\mathcal{M}, t)} \Rightarrow \forall \mathcal{N} \in \mathfrak{M} [\ \mathcal{M}_{\geq t} = \mathcal{N}_{\geq t} \Rightarrow \exists t1 \leq t\ \ \psi_{P(\mathcal{N}, t1)}]]$

The future formula $\psi_{F(\mathcal{M}, t)}$ is a *sufficient future interaction grounding for* $\varphi_{(\mathcal{M}, t)}$ if*:*

$\forall \mathcal{M} \in \mathfrak{M}\ \forall t [\ \forall \mathcal{N} \in \mathfrak{M} [\ \mathcal{M}_{\leq t} = \mathcal{N}_{\leq t} \Rightarrow \exists t1 \geq t\ \ \psi_{F(\mathcal{N}, t1)}\ ]\ \ \Rightarrow\ \ \exists t2 \leq t\ \varphi_{(\mathcal{M}, t2)}\ ]$

The future formula $\psi_{F(\mathcal{M}, t)}$ is a *necessary future interaction grounding for* $\varphi_{(\mathcal{M}, t)}$ if*:*

$\forall \mathcal{M} \in \mathfrak{M}\ \forall t [\ \varphi_{(\mathcal{M}, t)} \Rightarrow\ \forall \mathcal{N} \in \mathfrak{M} [\ \mathcal{M}_{\leq t} = \mathcal{N}_{\leq t} \Rightarrow \exists t1 \geq t\ \ \psi_{F(\mathcal{N}, t1)}]]$

5. CONCLUSIONS AND FURTHER RESEARCH

In the discussion on representational content of mental states, often the argument is made that for most mental properties no satisfactory way can be found to relate them to the (physical) world state, and hence symbolic or logical means are of no use to describe cognitive phenomena (the symbol grounding problem). Alternatives put forward (cf. (Clapin et al., 2000)) include the dynamical systems approach, and the interactionist perspective; cf. (Port and van Gelder, 1995; Bickhard, 1993; Christensen and Hooker, 2000). In line with these, in this paper the dynamic and interactionist perspective is adopted. It is shown how, if an interactionist perspective is taken, logical means in the form of temporal languages and semantics can successfully be used to describe the dynamics of mental states and properties, in relation to the dynamics of the interaction with the external world. Using this temporal approach, mental states and properties get their semantics in a formal manner in the temporal traces describing past and future interaction with the external world, in accordance with what is proposed informally by, e.g., (Bickhard, 1993; Christensen and Hooker, 2000; Clark, 1997).

The major difference with the work as mentioned is that in our approach a formalisation is proposed. This throws a new light on the sometimes assumed symbolic versus dynamics controversy. It shows how symbolic means can be used to describe dynamics as well; dynamics as a variety of phenomena entails no commitment to either Dynamical Systems Theory (DST) or symbolic methods as means to describe it.

The approach presented here contributes on the one hand a solid foundation for perspectives on dynamics and interaction as occurring in the recent literature. On

the other hand, the use of the temporal trace language has a number of practical advantages as well. In the first place, it offers a well-defined language to formulate relevant dynamic relations in practical domains, with standard first order logic semantics. It has a high expressive power. For example, the possibility of explicit reference to *time points* and *time durations* enables modelling of the dynamics of real-time phenomena, such as sensory and neural activity patterns in relation to mental properties (cf. (Port and van Gelder, 1995)). Also difference and differential equations can be expressed (see the extended paper). These features go beyond the expressivity available in standard linear or branching time temporal logics.

An interesting challenge for the temporal and interactionist perspective presented here is found in work based on the dual-level hypothesis, expressing that cognitive processes can be modelled according to two levels: the conceptual level (e.g., based on a symbolic model) and the sub-conceptual level (e.g., based on a connectionist model); cf. (Sun, 2000). The dual-level hypothesis would suggest to obtain a more refined temporal description of the dynamics that takes into account three elements (and their dynamic interaction): conceptual level mental properties, sub-conceptual properties, and the environment, where the sub-conceptual properties in a sense mediate between the conceptual properties and the environment. As both symbolic models and DST-style models are expressible in our language, it might be expected that also combinations of such types of models (and their interaction) can be expressed. This is planned as one of the issues for further research.

ACKNOWLEDGEMENTS

The authors have benefit from discussions about the subject with Vera Stebletsova, Wieke de Vries and Wouter Wijngaards.

REFERENCES

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

Clapin, H., Staines, P., and 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.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Agai*n. MIT Press, 1997.

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

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

Sun, R. (2000). Symbol grounding: a new look at an old idea. *Philosophical Psychology*, 13, 2000, pp. 149-172.