Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Cognitive Systems Research 4 (2003) 137–155

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

A temporal-interactivist perspective on the dynamics of mental states

Action editor: Vasant Honavar

Catholijn M. Jonker[a], Jan Treur[a,b,*]

[a]Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[b]Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

Received 16 July 2002; accepted 10 November 2002

Abstract

This paper addresses the dynamics of mental states in relation to the dynamics of the interaction with the external world. It contributes a formalised temporal-interactivist approach to these dynamics based on temporal traces for semantics and on a temporal trace language as an expressive means to formulate dynamic properties. The approach provides a foundation for the dynamic and interactivist perspective on cognitive phenomena.

Keywords: Dynamics, Interactivist, Temporal; Mental state

1. Introduction

In recent literature in the area of cognitive science and the philosophy of the mind, cognitive functioning is studied from a dynamic and interactivist perspective (Bickhard, 1993, 2000; Clark, 1997, 1999; Kelso, 1995; Port & Van Gelder, 1995; Clapin, Staines, & Slezak, 2000; Christensen & 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 system 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*

*Corresponding author. Present address: Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Tel.: +31-20-444-7763; fax: +31-20-444-7653.

E-mail addresses: jonker@cs.vu.nl (C.M. Jonker), http://www.cs.vu.nl/~jonker (C.M. Jonker), treur@cs.vu.nl (J. Treur), http://www.cs.vu.nl/~treur (J. Treur).

1389-0417/03/$ – see front matter
doi:10.1016/S1389-0417(02)00103-1

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. No formalisation is proposed in the recent literature on the interactivist perspective on cognition such as Bickhard (1993, 2000) and Christensen and Hooker (2000). In literature such as Port and Van Gelder (1995) on the dynamical systems approach, modelling techniques based on algebraic and difference or differential equations between continuous numerical variables are commonly used.

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 also depend on the environment's dynamics?
- How do future traces relate to mental states?
- How does the notion of functional role of a mental state relate to an interactivist perspective of a mental state?

To answer these questions, the temporal aspect of the dynamics of mental states and the interaction with the environment are studied. In this paper formalisation of the dynamics and interaction is proposed on the basis of formally defined *traces* and an expressive *temporal trace language*. The temporal trace language is used to formulate *dynamic properties* of these traces. The interaction with the environment can be either an ongoing process or a terminating process. The approach covers both cases.

The approach follows the view of Kim (1996) that the specification of a dynamic property of traces can be seen as a *temporal representation* or *temporal relational specification* of the internal state property. Kim views this as a way to account for a broad or wide content of mental properties:

''*The third possibility is to consider beliefs to be wholly internal to the subjects who have them but consider their contents as giving relational specifications of the beliefs. On this view, beliefs may be neural states or other types of physical states of organisms and systems to which they are attributed. Contents, then, are viewed as ways of specifying these inner states*; *wide contents, then, are specifications in terms of, or under the constraints of, factors and conditions external to the subject, both physical and social, both current and historical.* (. . .) *These properties are intrinsic, but their specifications or representations are extrinsic and relational, involving relationships to other things and properties in the world. It may well be that the availability of such extrinsic representations are essential to the utility of these properties in the formulation of scientific laws and explanations.* (. . .) . . . *in attributing to persons beliefs with wide content, we use propositions, or content sentences, to represent them, and these propositions* (*often*) *involve relations to things outside the persons. When we say that Jones believes that water is wet, we are using the content sentence 'Water is wet' to specify this belief, and the appropriateness of this sentence as a specification of the belief depends on Jones' relationship, past and present, to her environment.* (. . .) *The approach we have just sketched has much to recommend itself over the other two. It locates beliefs and other intentional states squarely within the subjects*; *they are internal states of the persons holding them, not something that somehow extrudes from them. This is a more elegant metaphysical picture than its alternatives. What is 'wide' about these states is their specifications or descriptions, not the states themselves.*'' (*Kim*, 1996, *pp.* 200–202).

The paper is organised as follows. First, as a basis, in Section 2 states and state properties are introduced. Next, Section 3 defines the notion of (temporal) trace and introduces the temporal trace language TTL with which dynamic properties can be expressed. In Section 4 internal states and internal state properties are formally related to sets of interaction traces to obtain their representational content or semantics. Section 5 addresses how sets of traces can be characterised by dynamic properties expressed in the temporal trace language. Criteria are identified and formalised that express when a dynamic property specification defines a class of interaction traces that can be related to a specific internal state property.

Such a dynamic property specification can be viewed as a temporal relational specification of the internal state property; see the quotation above from Kim (1996), pp. 200–202.

Section 6 addresses relational specifications of internal states, whereas Section 7 addresses external attribution of mental states. In Section 8 the issues are illustrated for the case of trust dynamics. Section 9 discusses the notion of a temporal-interactivist explanation. Section 10 shows that the temporal trace language is powerful enough to formalise modelling techniques often used within the dynamical systems approach, i.e. difference and differential equations. Section 11 positions the contribution of this paper with respect to other literature. The practical applicability of the work is discussed, as well as the supporting software environment that has been developed.

2. States and state properties

To describe dynamics, the notion of state is important. Dynamics will be described in the next section as evolution of states over time. The notion of state as used here is characterised on the basis of an ontology defining a set of physical and/or mental (state) properties (following, among others, Kim, 1998) that do or do not hold at a certain point in time. These properties are often called state properties to distinguish them from dynamic properties that relate different states over time. A specific state is characterised by dividing the set of state properties into those that hold, and those that do not hold, in the state. Examples of state properties are 'the agent is hungry', 'the agent has pain', 'the agent's body temperature is 37.5 °C', or 'the environmental temperature is 7 °C'. Real value assignments to variables are also considered as possible state property descriptions. For example, in a dynamical system approach based on variables x_1, x_2, x_3, x_4, that are related by differential equations over time, value assignments such as

$x_1 \leftarrow 0.06$
$x_2 \leftarrow 1.84$
$x_3 \leftarrow 3.36$
$x_4 \leftarrow -0.27$

are considered state descriptions. State properties are described by ontologies that define the concepts used.

2.1. Ontologies and state properties

To define states and state properties, the following different types of ontologies are used:

- IntOnt(A): to express *internal properties* of agent A
- InOnt(A): to express properties of the *input* of agent A
- OutOnt(A): to express properties of the *output* of the agent, and
- ExtOnt(A): to express properties of the *external* world (for A).

For example, the properties 'the agent A has pain', 'the agent's body temperature is 37.5 °C', may belong to IntOnt(A), whereas 'the environmental temperature is 7 °C', may belong to ExtOnt(A). The agent input ontology InOnt(A) defines properties for perception, the agent output ontology OutOnt(A) properties that indicate initiations of actions of A within the external world. The combination of InOnt(A) and OutOnt(A) is the *agent interaction ontology*, defined by InteractionOnt(A) = InOnt(A) ∪ OutOnt(A). The *overall ontology* for A is assumed to be the union of all ontologies mentioned above:

$$\mathrm{OvOnt(A)} = \mathrm{InOnt(A)} \cup \mathrm{IntOnt(A)} \cup \mathrm{OutOnt(A)} \cup \mathrm{ExtOnt(A)}.$$

As yet no distinction between physical and mental internal state properties is made; the formal framework introduced in subsequent sections does not assume such a distinction. If no confusion is expected about the agent to which ontologies refer, the reference to A is sometimes left out.

To formalise state property descriptions, ontologies are specified in a (many-sorted) first order logical format: an ontology is specified as a finite set of sorts, constants within these sorts, and relations and functions over these sorts (sometimes also called a signature). The example properties mentioned above then can be defined by nullary predicates (or proposition symbols) such as hungry, or pain, or by

using n-ary predicates (with $n \geq 1$) like has_temperature(body, 37.5), or has_value (x_1, 0.06), or has_temperature(environment, 7).

For a given ontology Ont, the propositional language signature consisting of all *state ground atoms* based on Ont is denoted by At(Ont). The *state properties* based on a certain ontology Ont are formalised by the propositions that can be made (using conjunction, negation, disjunction, implication) from the ground atoms and constitute the set SPROP(Ont).

2.2. Different types of states

(a) A *state* for ontology Ont is an assignment of truth-values {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*. For the agent STATES(IntOnt) is the set of all of its possible *internal states*. Moreover, STATES(InteractionOnt) denotes the set of all *interaction states*.

(b) The standard satisfaction relation $\models$ between states and state properties is used: $S \models p$ means that property p holds in state S. For a property p expressed in Ont, the set of states over Ont in which p holds (i.e. the S with $S \models p$) is denoted by STATES(Ont, p).

(c) For a state S over ontology Ont with subontology 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). The restriction operator serves as a form of projection of a combined state onto one of its parts.

3. Expressing dynamic properties

To describe the internal and external dynamics of the agent, explicit reference is made to time. Dynamic properties can be formulated that relate a state at one point in time to a state at another point in time. Some examples of dynamic properties of a certain agent are described below.

A simple example is the following dynamic property specification for belief creation based on observation:

Observational belief creation

'at any point in time t1 if the agent observes at t1 that it is raining, then there exists a point in time t2 after t1 such that at t2 the agent believes that it is raining'.

The persistence of a belief b over time can be specified by the dynamic property:

Belief persistence

'at any points in time t1 and t2 after t1, if the agent believes b at t1, then the agent will believe b at t2'.

An example of another type is trust monotonicity; this dynamic property specification about the dynamics of trust over time involves the comparison of two histories:

Trust monotonicity

'for any two possible histories, the better the agent's experiences with public transportation, the higher the agent's trust in public transportation'.

These examples were kept simple; they are just meant as illustrations. No attempt was made to make them as realistic as possible. To express such dynamic properties, and other, more sophisticated ones, the temporal trace language TTL is introduced.

3.1. Time frame and trace

First, in Section 3.1 the notion of trace is defined more explicitly. Next, in Section 3.2 the language to express dynamic properties is discussed.

(a) 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.

(b) A *trace* γ over an ontology Ont and time frame T is a time-indexed sequence of states

$$\gamma_t (t \in T)$$

in STATES(Ont), i.e. a mapping

$\gamma: T \rightarrow \mathrm{STATES}(\mathrm{Ont})$.

The set of all traces over ontology Ont is denoted by TRACES(Ont), i.e.
$\mathrm{TRACES}(\mathrm{Ont}) = \mathrm{STATES}(\mathrm{Ont})^T$.

(c) A *temporal domain description* W is a given set of traces over the overall ontology, i.e.

$W \subseteq \mathrm{TRACES}(\mathrm{OvOnt})$.

This set represents all possible developments over time (respecting the world's laws) of the part of the world considered in the application domain.

Different traces with respect to an agent A, can refer to different experiments with A involving different worlds, or different events generated in the world. For human beings one can think of a set of experiments in cognitive science, in which different experiments are not assumed to influence the behaviour of the agent. For software agents, it is possible to even erase the complete history (complete reset) and then activate the agent in a new world setting.

(d) Given a trace γ over the overall ontology OvOnt, the *input state* of an agent A at time point t, i.e. $\gamma_t \,|\mathrm{InOnt}(A)$, is also denoted by

$\mathrm{state}(\gamma, t, \mathrm{input}(A))$.

Analogously,

$\mathrm{state}(\gamma, t, \mathrm{output}(A))$

denotes the *output state* of the agent at time point t, and

$\mathrm{state}(\gamma, t, \mathrm{internal}(A))$

the *internal state*. The overall state of a system (agent and environment) at a certain moment, is denoted by $\mathrm{state}(\gamma, t)$. Again, if no confusion is expected about the particular agent, the reference to A can be left out.

(e) To focus on different aspects of the agent and time, traces can be *restricted* to specific *ontologies and time intervals*. The ontology parameter 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 specifies the part of the time frame of interest. The *restriction* $\gamma^{\mathrm{Ont}}_{\mathrm{Interval}}$ of a trace γ to time in Interval and information based on Ont is a mapping

$\gamma^{\mathrm{Ont}}_{\mathrm{Interval}}: \mathrm{Interval} \rightarrow \mathrm{STATES}(\mathrm{Ont})$,

defined by:

$\gamma^{\mathrm{Ont}}_{\mathrm{Interval}}(t) = \gamma(t)|\mathrm{Ont}$ if $t \in \mathrm{Interval}$.

For example, the *interaction trace* $\gamma^{\mathrm{InteractionOnt}}_{\leq t}$ denotes the restriction of γ to the past up to t and to interaction states.

3.2. Temporal trace language

To express dynamic properties in a precise manner a language is used in which explicit references can be made to time points and traces.

Comparable to the approach in situation calculus, the sorted predicate logic temporal trace language TTL is built on atoms referring to, for example, traces, time and state properties. For example, 'in the output state of A in trace γ at time t property p holds' is formalised by

$\mathrm{state}(\gamma, t, \mathrm{output}(A)) \models p$.

Here $\models$ is a predicate symbol in the language, usually used in infix notation, which is comparable to the Holds-predicate in situation calculus. Dynamic properties are expressed by temporal statements built using the usual logical connectives and quantification (for example, over traces, time and state properties). For example the following dynamic properties are expressed:

Observational belief creation

'in any trace, if at any point in time t1 the agent A observes that it is raining, then there exists a point in time t2 after t1 such that at t2 in the trace the agent A believes that it is raining'.

In formalised form:

$\forall \gamma \in W\ \forall t1$
$[\mathrm{state}(\gamma, t1, \mathrm{input}(A)) \models \mathrm{observation\_result}(\mathrm{itsraining})$
$\Rightarrow \exists t2 \geq t1\ \mathrm{state}(\gamma, t2, \mathrm{internal}(A)) \models \mathrm{belief}(\mathrm{itsraining})]$

Belief persistence

'in any trace, for any points in time t1 and t2 after t1, if the agent A has the belief b at t1 in the trace, then agent A has the belief b at t2 in this trace'.

In formalised form:

$$\forall \gamma \in W\ \forall t1, t2$$
$$[state(\gamma, t1, internal) \models b\ \&\ t1 \leq t2$$
$$\Rightarrow state(\gamma, t2, internal) \models b]$$

Trust monotonicity

'for any two traces $\gamma 1$ and $\gamma 2$, if at each time point t the agent A's experience with public transportation in $\gamma 2$ at t is at least as good as A's experience with public transportation in $\gamma 1$ at t, then in trace $\gamma 2$ at each point in time t, A's trust is at least as high as A's trust at t in trace $\gamma 1$'.

In formalised form:

$$\forall \gamma 1, \gamma 2 \in W$$
$$[\forall t\ [state(\gamma 1, t, input(A)) \models has\_value(experience, v1)\ \&$$
$$state(\gamma 2, t, input(A)) \models has\_value(experience, v2)$$
$$\Rightarrow v1 \leq v2]$$
$$\Rightarrow$$
$$\forall t\ [state(\gamma 1, t, internal(A)) \models has\_value(trust, w1)\ \&$$
$$state(\gamma 2, t, internal(A)) \models has\_value(trust, w2)$$
$$\Rightarrow w1 \leq w2]]$$

The set TFOR(Ont) is the set of all *temporal statements* or *temporal formulations* that only make use of ontology Ont. We allow additional language elements as abbreviations of statements of the temporal trace language. A *past statement* for γ and t is a temporal statement $\psi(\gamma, t)$ 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 statement. The set of past statements over ontology Ont with respect to time point t is denoted by PFOR(Ont, t). Note that for any past statement $\psi(\gamma, t)$ the following holds:

if two traces $\gamma 1$, $\gamma 2$ are equal up to time point t, then:

$\psi(\gamma 1, t)$ holds if and only if $\psi(\gamma 2, t)$ holds.

Formally:

$$\forall\ \gamma 1, \gamma 1 \in W\ [\gamma 1_{\leq t} = \gamma 2_{\leq t} \Rightarrow [\psi(\gamma 1, t) \Leftrightarrow \psi(\gamma 2, t)]]$$

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

4. Internal states and interaction dynamics

As put forward in the Introduction, according to the interactivist view, a possible internal state "... serves to implicitly categorise together that class of environments that would yield that final state if interacted with" (Bickhard, 1993). Using our framework introduced in Sections 2 and 3 the *set of interaction histories* over ontology Ont (e.g. InteractionOnt, InOnt, or OutOnt) leading to internal property p, is defined as follows:

PTRACES(Ont, p) is the set of all traces over Ont up to some time point t, that are the restriction of an overall trace in which at time t internal state property p holds (Fig. 1).

This is formally defined by:

$$PTRACES(Ont, p) =$$
$$\{\gamma_{\leq t}^{Ont} \mid t \in T, \gamma \in W, state(\gamma, t, internal) \models p\}$$

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 interaction futures* over Ont allowed by p is defined as follows:

FTRACES(Ont, p) is the set of all traces over Ont starting at some time point t, that are the restriction of an overall trace in which at time t internal state property p holds.

This is formally defined by:

$$FTRACES(Ont, p) =$$
$$\{\gamma_{\geq t}^{Ont} \mid t \in T, \gamma \in W, state(\gamma, t, internal) \models p\}$$

Based on these formal definitions, the *representational content of an internal state property* p with respect to the interaction is defined as the pair of sets

$$\langle PTRACES(InteractionOnt, p),$$
$$FTRACES(InteractionOnt, p)\rangle$$

The concepts introduced are illustrated by an example of the internal state property s. This property is assumed to have relationships to the input state property injury (by a stinging wasp). An injury causes increased sensitivity, denoted by the internal state property s (condition 1), and it is the only

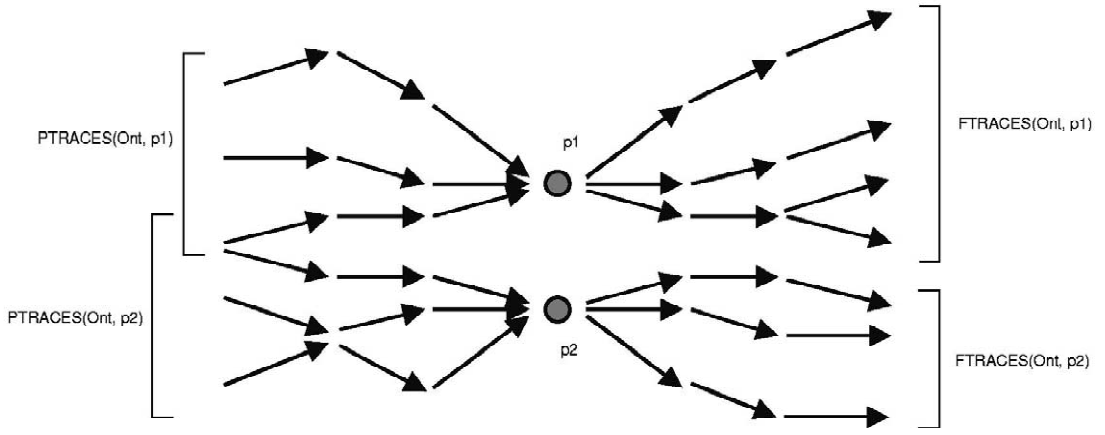

Fig. 1. Sets of past interaction traces PTRACES(Ont, p) and future interaction traces FTRACES(Ont, p) for p1 and p2.

possible cause (condition 2). The set of world traces W for this example reflects this in the sense that for any trace, always after injury occurs at the input, the internal state property s will occur further on in the trace (condition 1). Moreover, if s occurs in a trace, then earlier in the trace injury occurred at the input (condition 2). An example of an interaction history on the *input* of the agent leading to s, i.e. an element of PTRACES(InOnt, s), is the following (partially depicted) interaction trace:

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

In this trace injury takes place at time point t2. Another trace is:

t0. injury: false;
t1. injury: true;
t2. injury: false;
t3. injury: false;
t4. injury: false;

Here injury takes place at t1. 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 s. The amount of delay ≥ 0 taken into account is easily expressible in TTL by using the real numbers as time frame.

For the future perspective, the internal state property s is assumed to have relationships to the output state property move. The property s causes the action move depending on whether or not an environment object of the type that caused the injury (a wasp) stays close or returns (condition 1). It is assumed that these outputs are only generated if the internal state property s holds (condition 2). The set of world traces W reflects this in the sense that always after a time point where s occurs in the internal state, if later on at the input wasp_present occurs, i.e. the wasp is still there or returned, then this is followed by move at the output later on in the trace (within a certain response time d, which for simplicity will be left out). In other words, all wasps encountered in future will trigger an avoidance reaction. An example of a (partially depicted) interaction future allowed by s, i.e. an element of FTRACES(InteractionOnt, s), is as follows:

t3. wasp_present: false;
t4. wasp_present: true;
t5. move: true;

Here at t4 a wasp occurs; this is followed by a move action at t5. Another trace is:

t5. wasp_present: false;
t6. wasp_present: false;
t7. wasp_present: false;

In this trace no wasp occurs, and no move action is performed.

Consider a time point t and an overall trace γ

which future interaction part with respect to t is an element of the set of future interaction traces FTRACES(InteractionOnt, s) of s. In relation to the set of future interaction traces, the following question may arise: Is it always true that state(γ, t, internal) ⊨ s?

The answer to this question is: 'not necessarily'. If the future interaction part $\gamma_{\geq t}^{\text{InteractionOnt}}$ of an overall trace γ is in FTRACES(InteractionOnt, s), then this does not mean that state(γ, t, internal) ⊨ s. It only requires that an overall trace δ exists such that

$$\gamma_{\geq t}^{\text{InteractionOnt}} = \delta_{\geq t}^{\text{InteractionOnt}}$$

and

state(δ, t, internal) ⊨ s.

Furthermore, $\gamma_{\geq t}^{\text{InteractionOnt}} = \delta_{\geq t}^{\text{InteractionOnt}}$ does not mean that $\gamma_{\geq t}^{\text{IntOnt}} = \delta_{\geq t}^{\text{IntOnt}}$. Also, it might be that $\gamma_{<t}^{\text{InteractionOnt}} \neq \delta_{<t}^{\text{InteractionOnt}}$. Therefore, it is possible that in γ in the internal state no s occurs at time t, whereas in δ it does. So it is possible that after time t in γ no wasp ever gets close to the agent and still γ's interaction part from t onwards is an element of the set of future interaction traces. Concluding, traces like γ and time points t exist such that at time t the internal property s does not hold in t, but whose interaction behaviour from t onwards is indistinguishable from interaction behaviour in traces for which internal state s does hold at time t.

This discussion can be viewed as an illustration of the claim by Clark (1997, 1999): ''putting brain, body and world together again''. It is essential to consider overall traces in which the mental states, the world states, and the interactions between the two are covered. Without having an overall trace as a basis, it is quite possible to isolate one of these aspects (for example, the interaction), and loose the connection to the other aspects (for example, the mental states).

5. Dynamic properties characterising past and future interaction traces

Until now the interaction histories and futures have been defined semantically, by set-theoretic means in the form of sets of past interaction traces and future interaction traces. A natural question is whether such sets of past interaction traces and future interaction traces can be characterised by dynamic properties, and if so, by which ones. Using the temporal trace language introduced in Section 3, it is shown how such sets of traces can be characterised by dynamic properties expressed as temporal statements over traces, in two different ways. First, in Section 5.1 an example is discussed. Next, in Section 5.2 the notion of temporal relational specification of an internal state property is defined to characterise the sets of past and future interaction traces of this internal state property. In Section 5.3 a generalisation is made in the sense that no internal state property is the point of departure, but a dynamic property. The notion of trace relational specification for such a dynamic property is defined as a more general way to characterise sets of past and future interaction traces.

In Section 6 the implications of the general notion for the case that an internal state property exists are discussed. In Section 7 the case of external attribution of mental properties on the basis of observed behaviour is addressed (without assuming an internal state property).

5.1. Dynamic properties for the example

In the example, under a zero delay assumption, the set of past interaction traces PTRACES(InOnt, s) for internal state property s is characterised by the following (rather simple) dynamic property:

A past interaction trace up to time point t is in the set of past interaction traces for internal state property s

if and only if

it is the restriction of an overall trace in which *there is some time point t1 earlier than t at which injury occurs at the input*.

Expressed formally:

$$\gamma_{\leq t}^{\text{InteractionOnt}} \in \text{PTRACES}(\text{InOnt}, s) \Leftrightarrow \psi_P(\gamma, t)$$

where $\psi_P(\gamma, t) \in$ PFOR(InOnt, t) is the past dynamic property

$$\exists t1 \leq t \; \text{state}(\gamma, t1, \text{input}(A)) \models \text{injury}$$

The dynamic property $\psi_P(\gamma, t)$ can be considered as

specifying how the internal state property s relates to external events distant in time and/or space. This is a way to account for broad or wide (representational) content of mental state properties: by a *temporal relational specification for the past* of the internal state property s (Kim, 1996, pp. 200–202; see also the quotation in the Introduction; Fig. 2).

If a fixed delay d is taken into account, the existential quantifier in $\psi_P(\gamma, t)$ has to be instantiated by $t-d$. In that case the following holds:

$$\gamma^{InteractionOnt}_{\leq t} \in PTRACES(InOnt, s) \Leftrightarrow \psi_P(\gamma, t-d)$$

where $\psi_P(\gamma, t-d) \in PFOR(InOnt, t)$ is the past dynamic property

$$state(\gamma, t-d, input(A)) \models injury$$

If a delay with some randomness between 0 and d is assumed, then $\psi_P(\gamma, t)$ has to be defined as

$$\exists d'(0 \leq d' \leq d)\ state(\gamma, t-d', input(A)) \models injury$$

to guarantee the implication $\Rightarrow$. However, the other implication then does not hold.

For the future direction, the set of future interaction traces FTRACES(InteractionOnt, s) for internal state property s can be characterised as follows:

A future interaction trace from time point t is in the set of future interaction traces for internal state property s

if and only if

it is the restriction of an overall trace in which *for every time point t1 later than t, if at t1 a wasp occurs, then at some point in time t2 after t1 the agent moves.*

Expressed formally:

$$\gamma^{InteractionOnt}_{\geq t} \in FTRACES(InteractionOnt, s) \Leftrightarrow \psi_F(\gamma, t)$$

where the future dynamic property $\psi_F(\gamma, t) \in FFOR(InteractionOnt, t)$ is:

$$\forall t1\ t[state(\gamma, t1, input) \models wasp\_present \Rightarrow \exists t2 \geq t1\ state(\gamma, t2, output) \models move]$$

The dynamic property $\psi_F(\gamma, t)$ can be considered as a *temporal relational specification for the future* of the internal state property s (Kim, 1996, pp. 200–202).

Also here zero or a fixed delay is assumed. Notice that due to the conditional in $\psi_F(\gamma, t)$, if traces occur where never the condition on wasp_present comes to hold, then the implication is trivially true. This shows that it is not always possible on the basis of one trace to conclude by the implication $\Leftarrow$ that there has been s. See also the discussion on how this relates to the view of Clark (1997, 1999) at the end of Section 4.

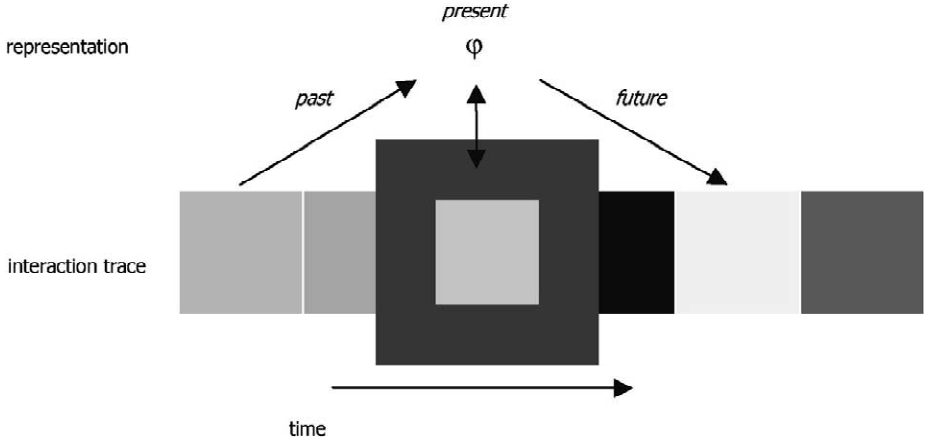

Fig. 2. Temporal relationships between internal states and interactions in past and future.

5.2. Temporal relational specifications

As a promising perspective for the discussion on broad or wide mental content of internal state properties, Kim (1996, pp. 200–202) puts forward the suggestion to consider wide content as a form of relational specification of an internal state property. The manner in which the example was analysed in Section 5.1 indeed follows this suggestion in a temporal sense. In accordance with interactivism, no reference to an independent external world is made, but only to interaction with such an external world; conform (Bickhard, 1993). This leads to a definition of wide content or representational content in the form of temporal relational specifications of an internal state property as follows.

Definition. (**Temporal relational specification**) A *temporal relational specification* of internal state property p is a pair of dynamic properties

$$\langle \psi_P(\gamma, t), \psi_F(\gamma, t) \rangle$$

with

$$\psi_P(\gamma, t) \in \mathrm{PFOR}(\mathrm{InteractionOnt}, t)$$

a past dynamic property and

$$\psi_F(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$$

a future dynamic property, such that the following hold:

(i) A past interaction trace up to time point t is in the set of past interaction traces for internal state property p if and only if it is the restriction of an overall trace γ for which $\psi_P(\gamma, t)$ holds.

Formally: for all overall traces γ and time points t it holds:

$$\gamma^{\mathrm{InteractionOnt}}_{\leq t} \in \mathrm{PTRACES}(\mathrm{InteractionOnt}, p) \Leftrightarrow \psi_P(\gamma, t)$$

(ii) A future interaction trace from time point t is in the set of future interaction traces for internal state property p if and only if it is the restriction of an overall trace γ for which $\psi_F(\gamma, t)$ holds.

Formally: for all overall traces γ and time points t it holds:

$$\gamma^{\mathrm{InteractionOnt}}_{\geq t} \in \mathrm{FTRACES}(\mathrm{InteractionOnt}, p) \Leftrightarrow \psi_F(\gamma, t)$$

The example illustrates that temporal relational specifications of an internal state property (in the sense of the past and future traces sets), depends on the assumption (1) that there is a fixed delay, and (2) that an internal state property exists for the considered notion (s). For a deterministic mathematical modelling approach, assumption (1) is customary (although not quite desirable), but if it can be weakened, this would be preferable. Assumption (2) 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 (2) would be artificial.

Below, in Section 5.3 it is shown how these assumptions can be avoided by involving a slightly more complex type of characterisation, based on mutual comparison of traces. By quantification over possible traces, this more sophisticated approach also gives more direct 'if and only if' correspondences than is possible in the case of one trace.

5.3. Trace relational specifications

Avoiding the assumptions discussed above, the following notions of relational specification are introduced. Notice that, compared to a temporal relational specification, instead of an internal state property p that is to hold at a certain time point, the more general condition that a dynamic property $\varphi(\gamma, t)$ holds is taken.

Definition. (**Trace relational specification**) Let $\psi_P(\gamma, t) \in \mathrm{PFOR}(\mathrm{InteractionOnt}, t)$ be a past dynamic property and $\psi_F(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$ a future dynamic property over the interaction ontology. Moreover, let Ont be a given ontology (e.g. the internal ontology), and $\varphi(\gamma, t) \in \mathrm{TFOR}(\mathrm{Ont})$ a temporal statement over Ont.

(a) The future dynamic property $\psi_F(\gamma, t)$ is a *sufficient future trace relational specification for* $\varphi(\gamma, t)$ if and only if the following holds for all traces γ and time points t:

if for any trace χ that coincides with γ in its past up to t, there is a time point $t1$ after t such that the

dynamic property $\psi_F(\chi, t1)$ holds, then a time point $t2$ before t exists such that property $\varphi(\gamma, t2)$ holds.

Formally:

$$\forall \gamma \in W \, \forall t \\ [\forall \chi \in W \, [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t \, \psi_F(\chi, t1)] \\ \Rightarrow \exists t2 \leq t \, \varphi(\gamma, t2)]$$

The future dynamic property $\psi_F(\gamma, t)$ is a *necessary future relational specification for* $\varphi(\gamma, t)$ if and only if the following holds for all traces γ and time points t:

if the property $\varphi(\gamma, t)$ holds, then for any trace χ that coincides with γ in its past up to t, there is a time point $t1$ after t such that the dynamic property $\psi_P(\chi, t1)$ holds.

Formally:

$$\forall \gamma \in W \, \forall t \\ [\varphi(\gamma, t) \Rightarrow \forall \chi \in W \, [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t \, \psi_F(\chi, t1)]]$$

(b) The past dynamic property $\psi_P(\gamma, t)$ is a *sufficient past trace relational specification for* $\varphi(\gamma, t)$ if and only if the following holds for all traces γ and time points t:

if for any trace χ that coincides with γ in its future starting at t, there is a time point $t1$ before t such that the dynamic property $\psi_P(\chi, t1)$ holds, then a time point $t2$ after t exists such that property $\varphi(\gamma, t2)$ holds.

Formally:

$$\forall \gamma \in W \, \forall t \\ [\forall \chi \in W \, [\gamma_{\geq t} = \chi_{\geq t} \Rightarrow \exists t1 \leq t \, \psi_P(\chi, t1)] \\ \Rightarrow \exists t2 \geq t \, \varphi(\gamma, t2)]$$

The past dynamic property $\psi_P(\gamma, t)$ is a *necessary past trace relational specification for* $\varphi(\gamma, t)$ if and only if the following holds for all traces γ and time points t:

if the property $\varphi(\gamma, t)$ holds, then for any trace χ that coincides with γ in its future starting at t, there is a time point $t1$ before t such that the dynamic property $\psi_P(\chi, t1)$ holds.

Formally:

$$\forall \gamma \in W \, \forall t \\ [\varphi(\gamma, t) \Rightarrow \forall \chi \in W [\gamma_{\geq t} = \chi_{\geq t} \Rightarrow \exists t1 \leq t \, \psi_P(\chi, t1)]]$$

To explain these rather abstract notions of relational specification, in the following two subsections they are instantiated to special cases: internal state properties, and externally attributed mental properties. Note that requiring all four conditions may be a quite strong demand. In many cases the sufficient past trace relational specification and necessary future trace relational specification conditions will already serve the purposes. They already define the path from the past via the present time point t to the future. However, the other two relational specifications may play a role if a form of closure assumption is made, namely that the *only* way of obtaining the future behaviour is the specified way.

6. Relational specifications of internal states

In this section the notions introduced in Section 5.3 are applied to a specific choice for the ontology Ont and the dynamic property $\varphi(\gamma, t) \in$ TFOR(Ont). The ontology IntOnt is chosen for Ont, and the statement state$(\gamma, t,$ internal$) \models p$ for some internal state property p is chosen for $\varphi(\gamma, t)$. This leads to the following definitions. Let $\psi_P(\gamma, t) \in$ PFOR(InteractionOnt, t) be a past dynamic property and $\psi_F(\gamma, t) \in$ FFOR(InteractionOnt, t) a future dynamic property over the interaction ontology.

Definition. The internal state property p has an *external trace relational specification* or *temporal representation* or *interaction grounding* given by the two dynamic properties $\psi_P(\gamma, t)$ and $\psi_F(\gamma, t)$ if the following conditions are fulfilled:

Sufficient future trace relational specification:

$$\forall \gamma \in W \, \forall t \, [\forall \chi \in W [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t \, \psi_F(\chi, t1)] \Rightarrow \\ \exists t2 \leq t \, \text{state}(\gamma, t2, \text{internal}) \models p]$$

Necessary future trace relational specification:

$$\forall \gamma \in W \, \forall t \, [\text{state}(\gamma, t, \text{internal}) \models p \Rightarrow \\ \forall \chi \in W \, [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t \, \psi_F(\chi, t1)]]$$

Sufficient past trace relational specification:

$$\forall \gamma \in W \, \forall t [\forall \chi \in W [\gamma_{\geq t} = \chi_{\geq t} \Rightarrow \exists t1 \leq t \, \psi_P(\chi, t1)] \Rightarrow \\ \exists t2 \geq t \, \text{state}(\gamma, t2, \text{internal}) \models p]$$

Necessary past trace relational specification:

$\forall\gamma \in W\, \forall t[state(\gamma, t, internal) \models p \Rightarrow$
$\forall\chi \in W[\gamma_{\geq t} = \chi_{\geq t} \Rightarrow \exists t1 \leq t\, \psi_P(\chi, t1)]]$

The dynamic properties $\psi_P(\gamma, t)$ and $\psi_F(\gamma, t)$ are considered as an explicit definition of the *representational content* of the internal state property p in terms of past and future interactions.

In an instantiated form (i.e. with particular instances of $\psi_P(\gamma, t)$ and $\psi_F(\gamma, t)$ substituted), the conditions above obtain temporal statements (dynamic conditions) that guarantee that everything functions well: *proper functioning axioms* for the internal state property p. This is a generalisation of the notion of functional role specification of an internal state property (Kim, 1996). This can be illustrated for the wasp example; the instances for the two (past and future) dynamic properties are given in Section 4. A more extensive example is addressed in Section 8: the internal state property trust in relation to a history of experiences.

The internal state property p in the conditions above can also be taken not as a specific property, but left unspecified. Then it functions as a variable that can be existentially quantified to express that an instantiation exists such that the conditions hold. As a special case, in this existentially quantified form the conditions can express the functional role of a mental property as a *second order property* over physical properties (Kim, 1998, pp. 19–20):

> ''*Functionalism takes mental properties and kinds as functional properties, properties specified in terms of their roles as causal intermediaries between sensory inputs and behavioural outputs, and the physicalist form of functionalism takes physical properties as the only potential occupants, or 'realizers', of these causal roles. To use a stock example, for an organism to be in pain is for it to be in some internal state that is typically caused by tissue damage, and that typically causes groans, winces, and other characteristic pain behaviour. In this sense being in pain is said to be a second-order property: for a system x to have this property is for x to have some first order property P that satisfies a certain condition D, where in the present case D specifies that P has pain's typical causes and typical effects. More generally, we can explain the idea of a second-order property in the following way. Let B be a set of properties; these are our first-order (or 'base') properties.* (. . .) *We then have this*:
>
> *F is a second-order property over set B of base (or first-order) properties iff F is the property of having some property P in B such that D(P) where D specifies a condition on members of B. Second-order properties therefore are second-order in that they are generated by quantification-existential quantification in the present case-over the base properties. We may call the base properties satisfying condition D the realizers of second-order property F.*''

This means that if we denote (the conjunction of) the four conditions expressed above by $D(p)$, then within the temporal trace language $\exists p D(p)$ is the formalisation of the second order property pointed out informally or semi-formally by Kim. In this form the conditions state that a physical realisation of the mental property exists, satisfying the functional role attributed to the mental property. The conditions serve as a specification of this functional role, in a generalised form.

7. External attribution of a mental property

The notion of trace relational specification offers a possibility to define when some mental state property can be attributed externally (on the basis of externally observable behaviour only), without making any commitment to actual internal states of the agent. The idea is to choose the ontology InteractionOnt for Ont, and the past dynamic property $\psi_P(\gamma, t) \in$ PFOR(InteractionOnt) for $\varphi(\gamma, t)$. On the basis of this choice it can be verified immediately that the sufficiency and necessity conditions for past interaction are trivially fulfilled. What remain are the future interaction conditions. This obtains the following definition;

Let $\psi_F(\gamma, t) \in$ FFOR(InteractionOnt, t) be a future dynamic property over the interaction ontology. A past dynamic property $\psi_P(\gamma, t) \in$ PFOR(InteractionOnt) is called a *historical temporal representation* or *past interaction grounding* for

an attributed mental property with *future interaction grounding* $\psi_F(\gamma, t)$ if and only if the following conditions are fulfilled:

Sufficiency condition:

$$\forall \gamma \in W\, \forall t$$
$$[\forall \chi \in W\, [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t\, \psi_F(\chi, t1)]$$
$$\Rightarrow \exists t2 \leq t\, \psi_P(\gamma, t2)]$$

Necessity condition:

$$\forall \gamma \in W\, \forall t\, [\psi_P(\gamma, t)$$
$$\Rightarrow \forall \chi \in W[\gamma_{\leq t} = \chi_{\leq t} \Rightarrow \exists t1 \geq t\, \psi_F(\chi, t1)]]$$

This definition can be illustrated for the wasp example (assuming no internal state property for s).

8. Example models of trust dynamics

In this section discrete and continuous models for trust dynamics depending on experiences are addressed. Trust is a non-trivial mental state property in the sense that it is related to a whole history of experiences, and not to only the most recent experience; it is a kind of cumulative mental state property.

8.1. Discrete trust dynamics

To illustrate the temporal-interactivist approach for a less simple example, a model for trust dynamics (i.e. trust states in relation to histories of positive or negative experiences) is addressed, adopted from Jonker and Treur (1999). In this model trust (e.g. in somebody selling special fruit offers) has three possible states (distrust, indifferent, trust). To keep complexity limited, only the current experience and the experiences two steps back in history are taken into account to determine a trust state at time point t, according to Table 1.

Table 1
Example: discrete model of trust dynamics

Experience histories			Trust
t−2	t−1	t	t
+	+	+	Trust
+	+	−	Indifferent
+	−	+	Indifferent
+	−	−	Distrust
−	+	+	Trust
−	+	−	Distrust
−	−	+	Indifferent
−	−	−	Distrust

Future behaviour concerns whether or not to buy special fruit offers from this person. The following past dynamic properties serve as *past temporal relational specification* of the different trust states:

(1) *trust state* trust.

A past temporal relational specification of the trust state trust is the past dynamic property $\psi_1(\gamma, t) \in$ PFOR(InOnt, t) defined by

$$\text{state}(\gamma, t, \text{input}) \vDash \text{pos\_exp} \wedge$$
$$\text{state}(\gamma, t-1, \text{input}) \vDash \text{pos\_exp}$$

(2) *trust state* indifferent.

A past temporal relational specification of the trust state indifferent is the past dynamic property $\psi_2(\gamma, t) \in$ PFOR(InOnt, t) defined by

$$[\text{state}(\gamma, t, \text{input}) \vDash \text{neg\_exp} \wedge$$
$$\text{state}(\gamma, t-1, \text{input}) \vDash \text{pos\_exp} \wedge$$
$$\text{state}(\gamma, t-2, \text{input}) \vDash \text{pos\_exp}]$$
$$\vee$$
$$[\text{state}(\gamma, t, \text{input}) \vDash \text{pos\_exp} \wedge$$
$$\text{state}(\gamma, t-1, \text{input}) \vDash \text{neg\_exp}]$$

(3) *trust state* distrust.

A past temporal relational specification of the trust state distrust is the past dynamic property $\psi_3(\gamma, t) \in$ PFOR(InOnt, t) defined by

$$\text{state}(\gamma, t, \text{input}) \vDash \text{neg\_exp} \wedge$$
$$[\text{state}(\gamma, t-1, \text{input}) \vDash \text{neg\_exp} \vee$$
$$\text{state}(\gamma, t-2, \text{input}) \vDash \text{neg\_exp}]$$

The following future dynamic properties serve as *future temporal relational specification* of the different trust states:

(4) *trust state* trust.

A future temporal relational specification of the trust state trust is the future dynamic property $\psi_4(\gamma, t) \in$ FFOR(InteractionOnt, t) defined by

$$[\text{state}(\gamma, t, \text{input}) \vDash \text{offer}$$
$$\Rightarrow \text{state}(\gamma, t+1, \text{output}) \vDash \text{accept}]$$

(5) *trust state* indifferent.

A future temporal relational specification of the trust state indifferent is the future dynamic property $\psi_5(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$ defined by

$$[\mathrm{state}(\gamma, t, \mathrm{input}) \models \mathrm{offer} \Rightarrow \mathrm{state}(\gamma, t+1, \mathrm{output}) \models \mathrm{accept}]$$

(6) *trust state* distrust.

A future temporal relational specification of the trust state distrust is the future dynamic property $\psi_6(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$ defined by

$$[\mathrm{state}(\gamma, t, \mathrm{input}) \models \mathrm{offer} \Rightarrow \mathrm{state}(\gamma, t+1, \mathrm{output}) \models \mathrm{reject}]$$

Note that the (immediate) future temporal relational specifications for trust and indifference are indistinguishable. The distinction can only be made taking longer time periods into account.

8.2. Continuous trust dynamics

This example focuses on an analysis of the mental state property trust over continuous time. It is assumed that trust in the weather forecast depends on one's experiences based on continuously monitoring the actual weather and comparing the observed weather with the predicted weather. For some of the patterns of behaviour, decisions may depend on your trust in the weather forecast. In particular, the decision to take an umbrella depends not only on the weather forecast, but also on your trust in the weather forecast. For example, when the weather forecast is not bad, but your trust is low, you still take an umbrella with you.

It is assumed that for each point in time your experience with the weather forecast is a modelled by value (real number) between -1 (negative experience) and 1 (positive experience). The accumulation of experiences in trust may be described by averaging the accumulation of the shaded area of the graph of experiences values over time, shown in Fig. 3. So, trust gives a kind of average of the experiences over time.

Based on this graph, for our example, trust is taken to be the real number indicating the shaded area divided by the length of the time interval, where the parts below the time axis count as negative. Within an overall trace the relation between trust

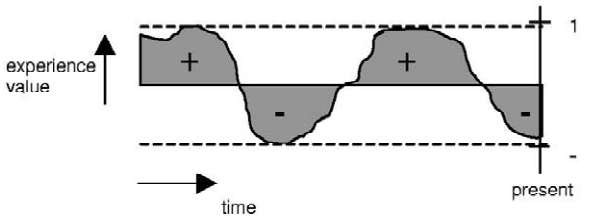

Fig. 3. Trust based on continuous experiences without decay.

value $\mathrm{tv}_\gamma(t)$ at a certain point in time $t>0$ and the interaction history can be modelled as the integral over time until t of the experience value $\mathrm{ev}_\gamma(t)$, i.e.

$$\mathrm{tv}_\gamma(t) = \int_0^t \mathrm{ev}_\gamma(u)\, du / t$$

where for a trace γ the functions tv, ev are defined by:

$$\mathrm{tv}_\gamma(t) = v \text{ iff } \mathrm{state}(\gamma, t, \mathrm{internal}) \models \mathrm{has\_value}(\mathrm{trust}, v)$$

$$\mathrm{ev}_\gamma(t) = w \text{ iff } \mathrm{state}(\gamma, t, \mathrm{input}) \models \mathrm{has\_value}(\mathrm{experience}, w)$$

Another way of modelling the same is by the differential equation

$$d\mathrm{tv}_\gamma(t)/dt = [\mathrm{ev}_\gamma(t) - \mathrm{tv}_\gamma(t)]/t$$

In a semantic manner, for a trust state w, the set of past interaction traces can be defined by

$$\mathrm{PTRACES}(\mathrm{InteractionOnt}, \mathrm{trust}(w)) = \{\gamma_{\le t}^{\mathrm{InteractionOnt}} \mid t \in T, \gamma \in W, \int_0^t \mathrm{ev}_\gamma(u)\, du = w\}$$

Another way of putting it is by the following characterisation:

$$\gamma_{\le t}^{\mathrm{InteractionOnt}} \in \mathrm{PTRACES}(\mathrm{InteractionOnt}, \mathrm{trust}(w)) \Leftrightarrow \int_0^t \mathrm{ev}_\gamma(u)\, du = w$$

The example shows that for continuous time models characterisations show up that are formulated in terms of integrals or differential equations. This

provides an interesting connection of the temporal-interactivist perspective to the dynamical systems theory as advocated, for example, in Kelso (1995) and Port and Van Gelder (1995). This connection will be addressed in more depth in Section 10.

In the above example, it may seem not realistic that experiences very far back in time count the same as recent experiences. In the accumulation of trust, experiences further back in time will often count less than more recent experiences, based on a kind of inflation rate, or increasing memory vagueness. Therefore a more realistic model may be obtained if for the agent it is assumed that the graph of experience values against time is modified as shown below to fit it between two curves that are closer to zero further back in time. Trust then is the accumulation of the areas in the graph below, where the parts below the time axis count negative.

The limiting curves can be based, for example, on an exponential function e^{at} with $a>0$ a real number related to the strength of the decay. Then the graph in Fig. 4 depicts the resulting function $ev_\gamma(t)\ e^{at}$.

Under these assumptions the relation between trust and experiences can be modelled by the integral

$$\int_0^t ev_\gamma(u)\, e^{au}\, du.a/(e^{at}-1)$$

where the factor $a/(e^{at}-1)$ is a normalisation factor to normalise trust in the interval $[-1, 1]$.

Equivalently, trust can be described by the following differential equation:

$$dtv_\gamma(t)/dt = [ev_\gamma(t) - tv_\gamma(t)].a\, e^{at}/(e^{at}-1)$$

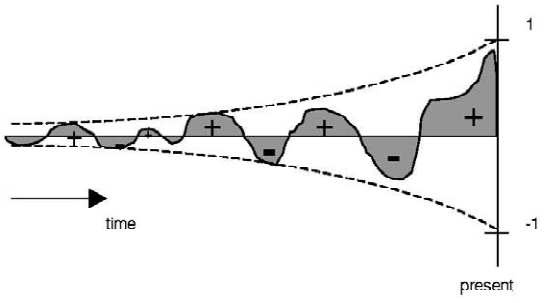

Fig. 4. Trust based on continuous experiences with decay.

The following characterisation follows for this case:

$$\gamma^{InteractionOnt}_{\le t} \in PTRACES(InteractionOnt, trust(w)) \Leftrightarrow \int_0^t ev_\gamma(u)\, e^{au}\, du.a/(e^{at}-1) = w$$

It is also possible to express this set in a syntactical manner in TTL. This requires that integral or differential equations are expressible in TTL, which will be addressed in Section 10.

9. Explanation from a temporal-interactivist perspective

In the trust example of Section 8.2, let us assume that the behaviour of an agent is as follows:

- If the weather forecast predicts fairly bad or seriously bad weather, you always take your umbrella with you
- If the weather forecast does not predict extremely nice weather, and your trust in the weather forecast is less than 0.2, then you take your umbrella with you.

If the agent behaviour is observed, and it turns out that the agent takes an umbrella with her while the current weather forecast is not bad. How can this behaviour be explained? A first step in an explanation is straightforward:

> '*The agent took an umbrella with her, because she does not trust the weather forecast*'.

This explanation is not different from an explanation from a functionalist perspective: trust has a (direct) functional role in taking umbrellas (Kim, 1996). But an iterated explanation will also ask: Why does the agent not trust the weather forecast? An explanation from a temporal-interactivist perspective will be:

> '*The agent does not trust the weather forecast because in her history she had a series of bad experiences that has accumulated in low trust*'.

This explanation does not fit well in the functionalist

perspective, as direct causes given for the current trust state will involve a more refined previous trust state, and so on. Such a process assumes a large number of refined trust states with trust dynamics described by discrete steps between these refined trust states, whereas in the temporal-interactivist approach a continuous temporal relationship can be used instead, and the trust states can be restricted to a modest number of them.

10. Dynamical systems approach

In Section 8, in some of the examples continuous relationships over time were encountered. These relationships were modelled semantically by differential equations, usually assumed to belong to the dynamical systems approach (DST), put forward, e.g. in Port and Van Gelder (1995). The question may arise whether or not such modelling techniques can be expressed in the temporal trace language TTL. In this section it is shown how modelling techniques used in the dynamical systems approach, such as difference and differential equations, can be represented in the temporal trace language. First the discrete case is considered. An example of an application is the study of the use of logistic and other difference equations to model growth (and in particular growth spurts) of various cognitive phenomena, e.g. the growth of a child's lexicon between 10 and 17 months (van Geert, 1991, 1995). The logistic difference equation used is:

$$L(n+1) = L(n)(1 + r - rL(n)/K)$$

Here r is the growth rate and K the carrying capacity. This equation can be expressed in our temporal trace language on the basis of a discrete time frame (e.g. the natural numbers) in a straightforward manner:

$$\forall \gamma \in W\ \forall t$$
$$\quad \mathrm{state}(\gamma, t, \mathrm{internal}) \models \mathrm{has\_value}(L, v) \Rightarrow$$
$$\quad \mathrm{state}(\gamma, t+1, \mathrm{internal}) \models \mathrm{has\_value}(L, v\,(1+r-rv/K))$$

The traces γ satisfying the above dynamic property are the solutions of the difference equation. Another illustration is the dynamical model for decision-making presented in Busemeyer and Townsend (1993) and Townsend and Busemeyer (1995). The core of their decision model for the dynamics of the preference P for an action is based on the differential equation

$$dP(t)/dt = -sP(t) + cV(t)$$

where s and c are constants and V is a given evaluation function. One straightforward option is to use a discrete time frame and model a discretised version of this differential equation along the lines discussed above. However, it is also possible to use the dense time frame of the real numbers, and to express the differential equation directly. To this end, the following relation is introduced, expressing that $x = dy/dt$:

$$\mathrm{is\_diff\_of}(\gamma, x, y):$$
$$\quad \forall t,w\ \forall \varepsilon{>}0\ \exists \delta{>}0\ \forall t',v,v'$$
$$\quad\quad 0 < \mathrm{dist}(t',t) < \delta\ \&\ \mathrm{state}(\gamma, t, \mathrm{internal}) \models \mathrm{has\_value}(x, w)$$
$$\quad\quad\quad \&\ \mathrm{state}(\gamma, t, \mathrm{internal}) \models \mathrm{has\_value}(y, v)$$
$$\quad\quad\quad \&\ \mathrm{state}(\gamma, t', \mathrm{internal}) \models \mathrm{has\_value}(y, v')$$
$$\quad\quad\quad \Rightarrow \mathrm{dist}((v'-v)/(t'-t), w) < \varepsilon$$

where $\mathrm{dist}(u,v)$ is defined as the absolute value of the difference, i.e. $u-v$ if this is ≥ 0, and $v-u$ otherwise. Using this, the differential equation can be expressed by:

$$\mathrm{is\_diff\_of}(\gamma, -sP + cV, P)$$

The traces γ for which this statement is true are (or include) solutions for the differential equation. Applied to the continuous trust dynamics addressed in Section 5, this can be expressed similarly by

$$\mathrm{is\_diff\_of}(\gamma, (\mathrm{experience} - \mathrm{trust})/t, \mathrm{trust})$$

Models consisting of combinations of difference or differential equations can be expressed in a similar manner. This shows how modelling techniques often used in the dynamical systems theory can be expressed in the temporal trace language TTL. In particular it shows that the relational specifications of trust states encountered in Section 8.2 can be expressed in TTL.

11. Discussion

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 (Clapin et al., 2000) include the dynamical systems approach, and the interactionist perspective (Port & van Gelder, 1995; Bickhard, 1993, 2000; Christensen & Hooker, 2000). In line with these, in this paper the dynamic and interactivist perspective is adopted.

It is shown how, if an interactivist 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, 2000), Christensen and Hooker (2000) and Clark (1997). In relation to the view on wide content of mental state properties as relational specifications, as put forward in Kim (1996, pp. 200–202), our approach gives a more detailed and formalised account—from a temporal and trace perspective—of these relational specifications.

The major difference with the work as mentioned is that in our approach a more detailed perspective and a formalisation is proposed. This throws 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.

Within the adopted agent-oriented modelling approach, recently developed formal conceptual modelling techniques and compositional verification and model checking techniques can be incorporated. Examples of such modelling techniques are process algebra; dynamic and temporal logic; event, situation and fluent calculus (Barringer, Fisher, Gabbay, Owens, & Reynolds, 1996; Eck et al., 2001; Fisher & Wooldridge, 1997; Reiter, 2001). These modelling techniques allow high-level expression of temporal relations, i.e. relations between a state of a process at one point in time, and states at other points in time. In addition, analysis techniques and tools, such as verification and model checking have progressed to a more mature status in recent years; e.g. Carnegie-Mellon's SMV, Cadence-SMV, and AT&T's SPIN (Clarke, Grumberg, & Peled, 2000; Manna & Pnueli, 1995; Stirling, 2001).

The approach presented here contributes on the one hand a solid foundation for perspectives on dynamics and interaction as occurring in recent literature. On the other hand, the use of the temporal trace language TTL 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 continuous real-time phenomena, such as sensory and neural activity patterns in relation to mental properties (Port & van Gelder, 1995). Also difference and differential equations can be expressed. These features go beyond the expressive power available in standard linear or branching time temporal logics.

The approach discussed above follows the standard view on calculus (based on epsilon-delta definitions). Recently, in Gamboa (2000) and Gamboa and Kaufmann (2001), an alternative approach, following the non-standard view (based on infinitesimals) has been presented for the integration of calculus within a logical (and theorem proving) framework. It may be the case, as claimed by some researchers, that for computational purposes the non-standard view has advantages. This will be an issue for further research.

Furthermore, the possibility to quantify over traces in TTL allows for specification of *more complex adaptive behaviours*. As within most temporal logics, reactiveness and pro-activeness properties are specified. In addition, in our language also properties expressing different types of adaptive behaviour can be expressed. For example a property such as 'exercise improves skill', which is a relative property in the sense that it involves the comparison of two alternatives for the history. Another property of this type is trust monotony: 'the better the experiences with something or someone, the higher the trust'. This type of relative property can be expressed in our

language, whereas in standard forms of temporal logic different alternative histories cannot be compared. Similarly, the kind of relative or comparative properties put forward in Jackson and Pettit (1990), such as 'the more south on the northern hemisphere, the higher the trees', as properties lacking an explanation in terms of a cause and its effects, can be expressed since our language allows comparison of different traces and different (local) restrictions within traces.

This possibility to define restrictions to *local languages for parts* of a system or the world is also an important feature. For example, the distinction between internal, external and input and output languages is crucial, and is supported by the language TTL, which also entails the possibility to quantify over system parts; this allows for specification of system modification over time. This possibility allows traces to be considered in which 'brain, body and world' are modelled in an integrative manner, and one of these aspects to be focussed on in the context of the overall trace (Clark, 1997, 1999).

Finally, since state properties are used as first class citizens in the temporal trace language, it is possible to explicitly refer to them, and to quantify over them, enabling the specification of what are sometimes called *second-order properties*, which are used in some of the philosophical literature (Kim, 1998) to express functional roles related to mental properties or states.

A practical advantage of the approach put forward is that based on the temporal trace language TTL a software environment has been developed consisting of two different tools. First, temporal statements expressed in the temporal trace language can be automatically checked against a set of traces, for example obtained from experiments or simulation. A software tool has been developed to support such a model checking process. In a second software tool developed, temporal trace statements of a specific 'leads to' format can be used to compute simulations, in the same spirit as executable temporal logic (Barringer et al., 1996).

The formalisation of the dynamics of mental state properties can also be applied to model and analyse the dynamics of reasoning processes. Work from this perspective has addressed the dynamics of reasoning about the control of reasoning (Treur, 2002), and defeasible reasoning processes from a temporal perspective (Engelfriet & Treur, 1995, 1998; Engelfriet, Marek, Treur, & Truszczinski, 2001) (see also Meyer and Treur, 2001). This work fully concentrates on the internal interaction and dynamics of mental states during a (defeasible) reasoning process; interaction with the external world is not addressed. Further work in this area will address the dynamics of practical reasoning processes based on assumptions, involving focussing of the reasoning on certain hypotheses that are assumed, predicting observable facts, interacting with the world to perform the observations, and evaluating the assumed focus hypotheses.

Acknowledgements

The authors have benefit from discussions about the subject with Vera Stebletsova, Wieke de Vries and Wouter Wijngaards. Lourens van der Mey and Wouter Wijngaards have contributed to the development of the software environment.

References

Barringer, H., Fisher, M., Gabbay, D., Owens, R., & Reynolds, M. (1996). *The imperative future: principles of executable temporal logic*. New York: John Wiley.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285–333.

Bickhard, M. H. (2000). Information and representation in autonomous agents. *Journal of Cognitive Systems Research*, *1*, 285–333.

Busemeyer, J., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Clapin, H., Staines, P., & Slezak, P. (Eds.), (2000). *Proceedings of the international conference on representation in mind: new theories of mental representation, 27–29th June 2000.* University of Sydney, Westport, CT, USA: Greenwood Publishers 2002.

Christensen, W. D., & Hooker, C. A. (2000). Representation and the meaning of life. In H. Clapin, P. Staines, & Slezak, P. (Eds.), *Proceedings of the international conference on representation in mind: new theories of mental representation, 27–29th June 2000.* University of Sydney (in press).

Clark, A. (1997). *Being there: putting brain, body and world together again*. Cambridge, MA, USA: MIT Press.

Clark, A. (1999). Where brain, body, and world collide. *Journal of Cognitive Systems Research*, *1*, 5–17.

Clarke, E. M., Grumberg, O., & Peled, D. A. (2000). *Model checking*. Cambridge, MA, USA: MIT Press.

Engelfriet, J., & Treur, J. (1995). Temporal theories of reasoning. *Journal of Applied Non-Classical Logics*, *5*, 239–261.

Engelfriet, J., & Treur, J. (1998). An interpretation of default logic in minimal temporal epistemic logic. *Journal of Logic, Language and Information*, *7*, 369–388.

Engelfriet, J., Marek, V. W., Treur, J., & Truszczinski, M. (2001). Default logic and specification of non-monotonic reasoning. *Journal of Experimental and Theoretical AI*, *13*, 99–112.

Fisher, M., & Wooldridge, M. (1997). On the formal specification and verification of multi-agent systems. *International Journal of Co-operative Information Systems*, *6*, 37–65.

Gamboa, R. (2000). Continuity and differentiability in ACL2. In Kaufmann, M., Manolios, P., & Moore, J. S. (Eds.), *Computer-aided reasoning: ACL2 case studies*. Dordrecht: Kluwer Academic Publishers, pp. 301–316.

Gamboa, R., & Kaufmann, M. (2001). Non-standard analysis in ACL2. *Journal of Automated Reasoning*, *27*, 323–351.

Jackson, F., & Pettit, P. (1990). Program explanation: a general perspective. *Analysis*, *50*, 107–117.

Jonker, C. M., & Treur, J. (1999). Formal analysis of models for the dynamics of trust based on experiences. In Garijo, F. J., & Boman, M. (Eds.), *Multi-agent system engineering, proceedings of the 9th European workshop on modelling autonomous agents in a multi-agent world, MAAMAW'99, Lecture notes in AI*, Vol. 1647. Berlin: Springer Verlag, pp. 221–232.

Kelso, J. A. S. (1995). *Dynamic patterns: the self-organisation of brain and behaviour*. Cambridge, MA: MIT Press.

Kim, J. (1996). *Philosophy of mind*. Westview.

Kim, J. (1998). *Mind in a physical world: an essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.

Manna, Z., & Pnueli, A. (1995). *Temporal verification of reactive systems: safety*. Berlin: Springer Verlag.

Meyer, J.-J., & Treur, J. (Vol. Eds.), (2001). *Dynamics and management of reasoning processes. Series in defeasible reasoning and uncertainty management systems* (Gabbay, D., Smets, Ph., Series Eds.), Dordrecht: Kluwer Academic Publishers.

Port, R. F., & Van Gelder, T. (Eds.), (1995). *Mind as motion: explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Reiter, R. (2001). *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. Cambridge, MA: MIT Press.

Stirling, C. (2001). *Modal and temporal properties of processes*. Berlin: Springer Verlag.

Townsend, J. T., & Busemeyer, J. (1995). Dynamic representation in decision making. In Port, R. F., & Van Gelder, T. (Eds.), *Mind as motion: explorations in the dynamics of cognition*. Cambridge, MA: MIT Press, pp. 101–120.

Treur, J. (2002). Formal semantics of meta-level architectures: dynamic control of reasoning. *International Journal of Intelligent Systems*, *17*, 545–568.

van Eck, P. A. T., Engelfriet, J., Fensel, D., van Harmelen, F., Venema, Y., & Willems, M. (2001). A survey of languages for specific dynamics: a knowledge engineering perspective. *IEEE Transactions on Knowledge and Data Engineering*, *13*(3), 462–496, May/June, 2001.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, *98*, 3–56.

van Geert, P. (1995). Growth dynamics in development. In Port, R. F., & Van Gelder, T. (Eds.), *Mind as motion: explorations in the dynamics of cognition*. Cambridge, MA: MIT Press, pp. 101–120.