A Dynamic Perspective on an Agent's Mental States and Interaction with its Environment

Catholijn M. Jonker[1] and Jan Treur[1,2]

[1] Vrije Universiteit Amsterdam
Department of Artificial Intelligence
De Boelelaan 1081a, 1081 HV Amsterdam
The Netherlands Tel. +31 20 444 77{43,63}

{jonker, treur}@cs.vu.nl

[2] Utrecht University
Department of Philosophy
Heidelberglaan 8, 3584 CS Utrecht
The Netherlands Tel. +31 30 253 2698

http://www.cs.vu.nl/~{jonker, treur}

ABSTRACT

This paper contributes a formalised foundation of the dynamics of an agent's mental states in relation to the dynamics of its interaction with the external world. The approach is based on trace semantics and provides a foundation for the dynamical and interactivist perspective on cognitive phenomena as known within Cognitive Science. A temporal trace language and a related software environment provide support for application.

Categories and Subject Descriptors
I.2.11 [**Computing Methodologies**]: Artificial Intelligence – Distributed AI – *Intelligent Agents*

Keywords
Agent, cognitive, dynamics, interactivist, philosophy, mind.

1. INTRODUCTION

In recent literature in the area of Cognitive Science and Philosophy of Mind, perspectives on cognitive functioning are proposed, where dynamics and interaction with the environment are central; e.g. [1], [2], [4], [5], [6], [7], [14], [18], [22]. For example, Bickhard [1] 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 are grounded in interaction histories on the one hand, and related to future interactions on the other hand. In the recent literature on the interactivist perspective on cognition such as [1], [2], [4], no formalisation is proposed. In literature such as [18] 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 interaction traces; how do they depend as well on the environment's future dynamics?
- How do they relate to mental states?
- How should the notion of functional role of a mental state be understood in an interactivist perspective?

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 or trajectories and an expressive temporal trace language to formulate dynamic properties of these traces. The approach covers both cases where termination in final states is assumed (as suggested in [1]) and interaction with the environment as an ongoing process.

First, as a basis, in Section 2 states and state properties are introduced. Next, in Section 3 the notion of trace and the temporal trace language TTL used are defined. In Section 4 it is shown how internal states and internal state properties can be formally related to sets of interaction traces to obtain their representational content or semantics. Section 5 addresses how these sets of traces can be characterised by dynamic properties, specified as formulae in TTL. Formal 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 relational specification or temporal grounding or temporal representation of this mental property. In Section 6 it is shown how the temporal trace language TTL covers formalisation of modelling techniques often used within the dynamical systems approach. Section 7 discusses the positioning of the contribution of this paper with respect to other literature and the practical applicability of the work, including a supporting software environment that has been developed.

AAMAS'02, July 15-19, 2002, Bologna, Italy.

2. STATES AND STATE PROPERTIES

Dynamics will be described in the next section as evolution of states over time. The notion of state as used here is defined on the basis of a fixed set of physical and/or mental properties (following, among others, [15]) that do or do not hold. A specific state is characterised by a distinction within this set of properties into the properties that hold in the state, and the (other) properties 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',
'the environmental temperature is 7° C'.

In particular, real value assignments to variables are considered as possible state property descriptions as well. 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 property descriptions. Properties are described by ontologies that define the concepts used.

2.1 State Ontologies and State Properties

To define states and state properties, the following different types of ontologies are used. An ontology for *internal properties* of the agent (IntOnt) for properties of the *input* (InOnt) and *output* (OutOnt) of the agent, and of the *external* world (ExtOnt). For example, the properties

'the agent has pain',
'the agent's body temperature is 37.5° C'

may belong to IntOnt, whereas

'the environmental temperature is 7° C',

may belong to ExtOnt. The agent input ontology InOnt defines properties for perception, the agent output ontology OutOnt properties that indicate initiations of actions of the agent. The combination of InOnt and OutOnt is the *agent interaction ontology*, defined by $\text{InteractionOnt} = \text{InOnt} \cup \text{OutOnt}$. The *overall ontology* is the union of all ontologies mentioned above:

$$\text{OvOnt} = \text{InOnt} \cup \text{IntOnt} \cup \text{OutOnt} \cup \text{ExtOnt}.$$

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.

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. The example properties mentioned above then can be defined by nullary predicates (or proposition symbols) such as hungry, or pain, or by using n-nary predicates (with $n \geq 1$) like

is_of_temperature(body, 37.5)
has_value (x_1, 0.06)
is_of_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 (finitary) conjunction, negation, disjunction, implication, from the ground atoms; they 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 (semantic) *satisfaction relation* $|=$ 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 sub-ontology Ont', a restriction of S to Ont' can be made, denoted by $S|\text{Ont'}$; this restriction is the member of STATES(Ont') defined by $S|\text{Ont'}(a) = S(a)$ if $a \in \text{At(Ont')}$. For example, if S is an overall state, i.e., a member of STATES(OvOnt), then the restriction of S to the internal atoms, $S|\text{IntOnt}$ 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. DYNAMICS IN A TEMPORAL TRACE LANGUAGE

To describe the internal and interaction dynamics of an agent, explicit reference is made to time in a formal manner.

3.1 Traces and Temporal Domain Description

a) A fixed *time frame* T is assumed which is linearly ordered. Depending on the application, for example 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).

b) A *trace* or *trajectory* γ over a state ontology Ont and time frame T is a mapping

$$\gamma : T \rightarrow \text{STATES(Ont)},$$

i.e., a sequence of states γ_t ($t \in T$) in STATES(Ont). The set of all traces over ontology Ont is denoted by TRACES(Ont), i.e., $\text{TRACES(Ont)} = \text{STATES(Ont)}^T$.

c) A *temporal domain description* W is a given set of traces over the overall ontology, i.e., $W \subseteq \text{TRACES(OvOnt)}$. This set represents all possible developments over time (respecting the world's laws) of the part of the world considered in the application domain.

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

state(γ, t, input).

Analogously, state(γ, t, output) denotes the output state of the agent at time point t, and state(γ, t, internal) the internal state. We can also refer to the overall state of a system (agent and environment) at a certain moment, denoted by state(γ, t).

e) To focus on different aspects of the agent and time, traces can be restricted to specific state 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* $\gamma_{\text{Interval}}^{\text{Ont}}$ of a trace γ to time in Interval and information based on Ont is a mapping

$$\gamma_{\text{Interval}}^{\text{Ont}}\text{: Interval} \to \text{STATES(Ont)},$$

defined by: $\gamma_{\text{Interval}}^{\text{Ont}}(t) = \gamma(t)|\text{Ont}$ if $t \in \text{Interval}$. For example, the *interaction trace* $\gamma_{\leq t}^{\text{InteractionOnt}}$ denotes the restriction of γ to the past up to t and to interaction atoms.

f) As in Section 2.2b), states within a trace at some point in time can be related to state properties via the (semantic) satisfaction relation |= between states and formulae. If $\varphi \in \text{SPROP(InOnt)}$, then $\text{state}(\gamma, t, \text{input}) \models \varphi$ denotes that φ is true in this state at time point t.

3.2 Temporal Trace Language

To express dynamic properties of traces the Temporal Trace Language TTL is used. Comparable to the approach in situation calculus, the sorted predicate logic language TTL is built on syntactic atoms referring to, e.g., traces, time and state properties, such as $\text{state}(\gamma, t, \text{output}) \models p$. Here |= is a (syntactic) predicate symbol in the language, comparable to the Holds-predicate in situation calculus. Note that this is in contrast with Section 2, where the same notation was used to denote the semantic relation. Using the same notation for both need not cause confusion: within TTL formulae always the syntactic symbol is meant.

Formulae in TTL 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 ontology Ont. We allow additional language elements as abbreviations of formulae of the temporal trace language. A *past formula* for γ and t is a temporal formula $\psi(\gamma, t)$ 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 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 $\psi(\gamma, t)$ it holds:

$$\forall \gamma, \chi \in W\ \forall t\ [\gamma_{\leq t} = \chi_{\leq t} \Rightarrow [\psi(\gamma, t) \Leftrightarrow \psi(\chi, t)]].$$

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 $s > t$.

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', cf. [1]. Using our framework introduced in Sections 2 and 3 the set of *interaction histories* or *past interaction traces* over ontology Ont leading to internal property p, is formally defined by

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

In addition, the way in which internal properties themselves lead to particular possible types of future interactions is also crucial for their meaning [1]. Therefore, for an internal state property p, and an ontology Ont, the set of *all future interaction traces* for t over Ont allowed by p is defined by:

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

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

The concepts introduced are illustrated by an example of the internal state property s (sensitivity for wasps). This property is assumed to have relationships to the input property injury, which causes sensitivity s (1), and is the only possible cause (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 (1). Moreover, if s occurs in a trace, then earlier in the trace injury occurred at the input (2). For the sake of simplicity, once the property s is there, it is assumed to persist over time. 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. input	injury: false
t1. input	injury: true
t2. input	injury: true

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. How much delay ≥ 0 is taken into account is

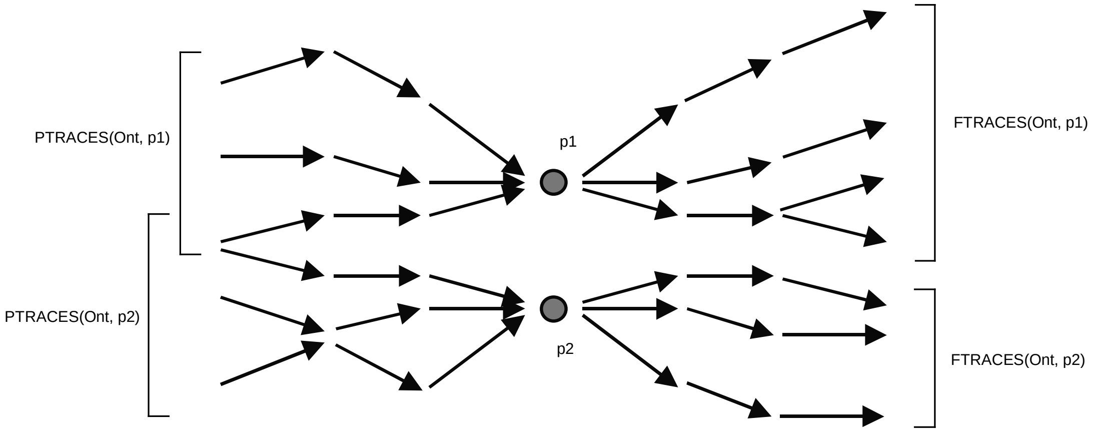

Figure 1. Sets of past interaction traces PTRACES(Ont, p) and future interaction traces FTRACES(Ont, p) for p1 and p2

easily expressible in the temporal approach introduced by taking the real numbers as time frame.

For the future perspective, the internal state property s is assumed to have relationships to the output property move. The property s causes the action move depending on whether or not a wasp stays close or returns (1). Moreover, it is assumed that these outputs are only generated systematically if the internal state property s holds (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, then this is followed by move at the output later on in the trace (within a certain maximal response 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 s, i.e., in FTRACES(InteractionOnt, s), is as follows:

t0. input	wasp_present: false
t1. input	wasp_present: true
t2. output	move: true

In relation to the set of future interaction traces, the following question may arise: for an overall trace γ with its future interaction part in the future set FTRACES(InteractionOnt, p) of p, is always state(γ , t, internal) |= p ? The answer on this question is: not necessarily; the future behaviour entailed by p may depend on a future condition in the environment. For example, if in γ no wasp is coming, then its interaction part may be in the future set, even if in the internal state no s occurs at time t in γ, because the interaction behaviour is indistinguishable from interaction behaviour in traces with internal state s at time t. However, if the future interaction part of an overall trace γ is in FTRACES(InteractionOnt, p), then at least an overall trace δ exists such that $\gamma_{\geq t}^{Ont} = \delta_{\geq t}^{Ont}$ and state(δ, t, internal) |= p. This can be viewed as an illustration of Clark [6], [7]'s claim '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 well 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 AS RELATIONAL SPECIFICATIONS

Until now the interaction histories and futures have been defined by semantic, set-theoretic means. However, using the temporal trace language TTL introduced in Section 3, sets of traces can be characterised in a syntactic manner by temporal formulae as well.

5.1 Temporal Relational Specifications

In the wasp example, under a zero delay assumption, the set PTRACES(InOnt, s) is characterised by:

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

where $\psi_P(\gamma, t) \in$ PFOR(InteractionOnt, t) is the past formula

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

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 socalled *broad* or *wide* (*representational*) *content* of mental state properties: by a *temporal relational specification* for the past of the internal state property s; cf. [15], p. 200-202

If a fixed delay d is taken into account, $\psi_P(\gamma, t)$ has to be replaced by $\psi_P(\gamma, t\text{-}d)$, to make the equivalence hold. If a delay with some randomness between 0 and d is assumed, then $\psi_P(\gamma, t)$ has to be replaced by $\exists d' \;\; (0 \leq d' \leq d) \wedge \psi_P(\gamma, t\text{-}d)$, to guarantee the implication $\Rightarrow$. However, the other implication then does not hold.

For the future direction, to characterise the set of traces FTRACES(InteractionOnt, s) in the form

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

a candidate formula $\psi_F(\gamma , t) \in$ FFOR(InteractionOnt, t) is:

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

As a promising perspective for the discussion on broad or wide mental content of internal state properties in [15], p. 200-202, the suggestion is put forward to consider wide content as a form of relational specification of an internal state property. The manner in which the example was analysed above indeed follows this suggestion in a temporal sense. The dynamic property $\psi_F(\gamma, t)$ can be considered as a *temporal relational specification for the future* of the internal state property s; cf. [15], p. 200-202. Also here zero or a fixed delay has to be assumed to make it work. This guarantees the implication $\Rightarrow$. However, 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 is or has been sensitivity.

In accordance with interactivism, no reference to an independent external world is made, but only to interaction with such an external world; cf [1]. 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

$$< \psi_P(\gamma, t) \;\; , \psi_F(\gamma, t) >$$

with $\psi_P(\gamma, t) \in$ PFOR(InteractionOnt, t) a dynamic past interaction property and $\psi_F(\gamma, t) \in$ FFOR(InteractionOnt, t) a dynamic future interaction 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:
$$\gamma_{\leq t}^{InteractionOnt} \in \text{PTRACES(InteractionOnt, p)} \quad \Leftrightarrow \quad \psi_P(\gamma, t)$$
(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:
$$\gamma_{\geq t}^{InteractionOnt} \in \text{FTRACES(InteractionOnt, p)} \quad \Leftrightarrow \quad \psi_F(\gamma, t)$$

In the example temporal relational specifications of the internal state property s (in the sense of the past and future interaction traces sets), depends on the assumption (1) fixed delay, and (2) 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 preferrable. 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 (to be) addressed, then assumption (2) would be artificial.

Below, In Section 5.2 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. In Section 5.3 the implications of the general notion for the case that an internal state property exists are discussed. In Section 5.4 the case of external attribution of mental properties on the basis of observed behaviour is addressed (without assuming an internal state property).

5.2 Trace Relational Specifications

Avoiding the assumptions discussed above, the following notions of trace relational specification are introduced. Instead of focusing on one trace, as in Section 5.1, a trace relational specification takes all possible continuation traces into account.

Definition (Trace Relational Specification)

Let $\psi_P(\gamma, t) \in \mathrm{PFOR}(\mathrm{InteractionOnt}, t)$ be a past formula and $\psi_F(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$ a future formula 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 formula over Ont.

The future formula $\psi_F(\gamma, t)$ is a *sufficient future interaction trace relational specification for* $\varphi(\gamma , t)$ if*:*

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

The future formula $\psi_F(\gamma, t)$ is a *necessary future interaction trace relational specification for* $\varphi(\gamma , t)$ if*:*

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

The past formula $\psi_P(\gamma, t)$ is a *sufficient past interaction trace relational specification for* $\varphi(\gamma, t)$ if*:*

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

The past formula $\psi_P(\gamma, t)$ is a *necessary past interaction trace relational specification for* $\varphi(\gamma , t)$ if*:*

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

To explain these rather abstract notions, in the following two subsections they are instantiated to special cases. Note that requiring all four conditions may be a quite strong demand. In many cases the sufficient past interaction trace relational specification and necessary future interaction trace relational specification conditions will already serve the purposes. They already define the path from history via present to future. However, the other two conditions 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.

5.3 Relational Specification of Internal States

In this subsection the notions introduced in Section 5.2 are applied to a specific choice for the ontology Ont and the temporal formula $\varphi(\gamma, t) \in \mathrm{TFOR}(\mathrm{Ont})$. The ontology IntOnt is chosen for Ont, and the formula $\mathrm{state}(\gamma, t, \mathrm{internal}) \models p$ for some internal state property p is chosen for $\varphi(\gamma, t)$. This leads to the following definition. Let $\psi_P(\gamma, t) \in \mathrm{PFOR}(\mathrm{InteractionOnt}, t)$ be a past formula and $\psi_F(\gamma, t) \in \mathrm{FFOR}(\mathrm{InteractionOnt}, t)$ a future formula over the interaction ontology. The internal state property p has a *trace relational specification* or an *external temporal representation* or *interaction grounding* given by the two formulae $\psi_P(\gamma, t)$ and $\psi_F(\gamma, t)$ if the following conditions are fulfilled:

Sufficiency condition for future interaction trace relational specification:

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

Necessity condition for future interaction trace relational specification:

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

Sufficiency condition for past interaction trace relational specification:

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

Necessity condition for past interaction trace relational specification:

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

The formulae $\psi_P(\gamma, t)$ 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 formulae (dynamic conditions) that guarantee that everything functions well: *proper functioning axioms* for the internal state property p. This can be illustrated for the wasp example; the instances for the two (past and future) formulae are given in Section 4.

The internal state property p in the conditions above can also be taken not 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; see, e.g., [16], 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.

5.4 Attribution of Mental Properties

The notion of relational specification offers a possibility to define when some mental 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 formula $\psi_P(\gamma, t) \in \text{PFOR(InteractionOnt)}$ 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 automatically fulfilled. What remain are the future interaction conditions. This obtains the following definition.

Let $\psi_F(\gamma,\ t) \in \text{FFOR(InteractionOnt,}\ t)$ be a future formula over the interaction ontology. A past formula $\psi_P(\gamma, t) \in \text{PFOR(InteractionOnt)}$ is called a *past interaction trace relational specification* or *historical temporal representation* or *past interaction grounding* for an attributed mental property with *future interaction trace relational specification* $\psi_F(\gamma,\ t)$ if the following conditions are fulfilled:

Sufficiency condition:

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

Necessity condition:

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

Also this definition can be illustrated for the wasp example (assuming no internal state property for sensitivity).

5.5 Example: Trust Dynamics

To illustrate the approach for a less simple example than the wasp example, a model for trust dynamics (based on positive (+) or negative (-) experiences) is addressed, adopted from [11]. In this model trust (in somebody selling cars) has three possible states (distrust, indifferent, trust). Only the current experience and the experiences two steps back in history are taken into account to determine trust, according to Table 1 below.

Table 1 Example 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 accept a very attractive car offer if put forward by this person. The following past formulae serve as *past interaction trace relational specification* of the different trust states:

(1) trust state trust

Past interaction trace relational specification of the trust state trust is the past formula $\psi_1(\gamma, t) \in \text{PFOR(InteractionOnt, t)}$ defined by

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

(2) trust state indifferent

Past interaction trace relational specification of the trust state indifferent is the past formula $\psi_2(\gamma, t) \in \text{PFOR(InteractionOnt, t)}$ defined by

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

(3) trust state distrust

Past interaction trace relational specification of the trust state distrust is the past formula $\psi_3(\gamma, t) \in \text{PFOR(InteractionOnt, t)}$ defined by

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

The following future formulae serve as *future interaction trace relational specification* of the different trust states:

(4) trust state trust

Future interaction trace relational specification of the trust state trust is the future formula $\psi_4(\gamma, t) \in \text{FFOR(InteractionOnt, t)}$ defined by

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

(5) trust state indifferent

Future interaction trace relational specification of the trust state indifferent is the future formula $\psi_5(\gamma,\ t) \in \text{FFOR(InteractionOnt,}\ t)$ defined by

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

(6) trust state distrust

Future interaction trace relational specification of the trust state distrust is the future formula $\psi_6(\gamma\ , t) \in \text{FFOR(InteractionOnt, t)}$ defined by

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

6. DYNAMICAL SYSTEMS APPROACH

In this section it is shown how modelling techniques used in the dynamical systems approach (DST) put forward, e.g., [18], 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 the lexicon between 10 and 17 months; cf. [9], [10]. The logistic difference equation used is:

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

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:

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

The traces γ satisfying the above formula are the solutions of the difference equation. Another illustration is the dynamical model for decision making presented in [3], [21]. The core of their decision model for the dynamics of the preference P for an action is based on the differential equation

$$dP(t)/dt = -s\ P(t)\ + c\ V(t)$$

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 $x = dy/dt$):

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

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:

$$\mathrm{is\_diff\_of}(\gamma\ , -\ s\ P\ \ + c\ V,\ P)$$

The traces γ for which this formula is true are (or include) solutions for the differential equation. 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 approach can be expressed in the temporal trace language.

7. 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; e.g., [20]). Alternatives put forward include the dynamical systems approach, and the interactivist perspective; cf. [18], [1], [2] [4]. 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., [1], [4], [6], [7].

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 in the first place a solid foundation for perspectives on dynamics and interaction as occurring in the recent literature. Additionally, 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 real-time phenomena, such as sensory and neural activity patterns in relation to mental properties (cf. [18]). Also difference and differential equations can be expressed. These features go beyond the expressivity available in standard linear or branching time temporal logics. Furthermore, the possibility to quantify over traces allows for specification in TTL of *more complex adaptive behaviours*. 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; for example, in LTL or Quantified LTL [19] no quantification over traces is possible.

For the temporal trace language TTL a software environment has been developed including an editor to specify dynamic properties and a (model) checker to check given traces against dynamic properties specified in TTL. As the current paper focuses on foundational issues, for reasons of space limitation details about this software enviroment have been left out.

The formalisation of the dynamics of mental state properties can also be applied to analyse the dynamics of reasoning processes. Work from this perspective has addressed defeasible reasoning processes from a temporal perspective: e.g., [8]; see also [17]. This work concentrates on the internal dynamics of mental states during a (defeasible) reasoning process; interaction with the external world is not addressed. Recent work has addressed the dynamics of practical reasoning processes based on multiple (e.g., geometric and arithmetic) representations: [12]. Further work addresses the dynamics of reasoning based on assumptions, involving focussing of the reasoning on certain hypotheses that are assumed, performing prediction of observable facts, having interaction with the world to perform the observations, and evaluating the assumed focus hypotheses. Furthermore, in [13] the dynamics of attributed beliefs, desires and intentions is addressed from a temporal perspective, as a continuation of Section 5.4 above.

Another interesting challenge for the temporal and interactivist 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. [20]. 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 is 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. Lourens van der Mey and Wouter Wijngaards have contributed to the development of the software environment.

REFERENCES

[1] Bickhard, M.H., Representational Content in Humans and Machines. Journal of Experimental and Theoretical Artificial Intelligence, 5, 1993, pp. 285-333.

[2] Bickhard, M.H., Information and representation in autonomous agents. Journal of Cognitive Systems Research, vol. 1, 2000, pp. 285-333.

[3] Busemeyer, J., and Townsend, J.T., Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, vol. 100, 1993, pp. 432-459.

[4] Clapin, H., Staines, P., and Slezak, P., 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.

[5] Clark, A., Being There: Putting Brain, Body and World Together Again. MIT Press, 1997.

[6] Clark, A., Where brain, body, and world collide. Journal of Cognitive Systems Research, 1, 1999, pp. 5-17.

[7] Christensen, W.D. and C.A. Hooker, Representation and the Meaning of Life. In: 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.

[8] Engelfriet, J., and Treur, J., Temporal Theories of Reasoning. Journal of Applied Non-Classical Logics, 5, 1995, pp. 239-261.

[9] Geert, P. van, A dynamic systems model of cognitive and language growth. Psychological Review, vol. 98, 1991, pp. 3-56.

[10] Geert, P. van, Growth Dynamics in Development. In: Port, R.F., Gelder, T. van (eds.), Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge, Mass., 1995, pp. 101-120.

[11] Jonker, C.M., and Treur, J., Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: F.J. Garijo, M. Boman (eds.), Multi-Agent System Engineering, Proc. of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99. Lecture Notes in AI, vol. 1647, Springer Verlag, 1999, pp. 221-232.

[12] Jonker, C.M., and Treur, J., Analysis of the Dynamics of Reasoning Using Multiple Representations. In: W.D. Gray and C.D. Schunn (eds.), Proceedings of the 24th Annual Conference of the Cognitive Science Society, CogSci 2002. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. In press, 2002.

[13] Jonker, C.M., Treur, J., and Vries, W. de, Temporal Requirements for Anticipatory Reasoning about Intentional Dynamics in Social Contexts. In: Y. Demazeau, F. Garijo (eds.), Multi-Agent System Organisations. Proc. of the 10th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'01, 2001.

[14] Kelso, J.A.S., Dynamic Patterns: the Self-Organisation of Brain and Behaviour. MIT Press, Cambridge, Mass, 1995.

[15] Kim, J., Philosophy of Mind. Westview press, 1996.

[16] Kim, J., Mind in a Physical world: an Essay on the Mind-Body Problem and Mental Causation. MIT Press, Cambridge, Mass, 1998.

[17] Meyer, J.-J., Ch., and Treur, J. (eds.), Dynamics and Management of Reasoning Processes. Series in Defeasible Reasoning and Uncertainty Management Systems (D. Gabbay, Ph. Smets, series eds.), Kluwer Academic Publishers, 2001.

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

[19] Sistla A.P. , M.Y. Vardi, and P. Wolper, The complementation Problem for Büchi Automata with Applications to Temporal Logic, Theoretical Computer Science, vol. 49, 1987, pp. 217-237.

[20] Sun, R., Symbol grounding: a new look at an old idea. Philosophical Psychology, 13, 2000, pp. 149-172.

[21] Townsend, J.T., and Busemeyer, J., Dynamic Representation in Decision Making. In: Port, R.F., Gelder, T. van (eds.), Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge, Mass., 1995, pp. 101-120.

[22] West, R.L., and Lebiere, C., Simple games as dynamic, coupled systems: randomness and other emergent properties. Journal of Cognitive Systems Research, 1, 2001, pp. 221-239.