# Temporal Analysis of the Dynamics of Beliefs, Desires, and Intentions[1]

**Catholijn M. Jonker, Jan Treur, and Wieke de Vries**

*Department of Artificial Intelligence, Vrije Universiteit Amsterdam* [2]
*Department of Philosophy, Utrecht University*

**Abstract** In this paper temporal relationships are expressed that provide an external temporal grounding of intentional notions. Justifying conditions are presented that formalise criteria that a (candidate) statement must satisfy in order to qualify as an external representation of a belief, desire or intention. Using these external representations, anticipatory reasoning about intentional dynamics can be performed.

Keywords: belief, desire, intention, temporal, dynamics, attribution

## 1 Introduction

As agent behaviour often goes beyond purely reactive behaviour, nontrivial means are needed to understandably describe and predict it. An attractive feature of intentional notions (cf. (Cohen and Levesque, 1990; Linder, Hoek, and Meyer, 1996; Rao and Georgeff, 1991)) to describe agent behaviour is that these notions offer a high level of abstraction and have intuitive connotations. As opposed to explanations from a direct physical perspective (the physical stance), in (Dennett, 1987, 1991) the *intentional stance* is put forward. Dennett emphasizes the advantage in tractability of intentional stance explanations for mental phenomena over physical stance explanations:

> 'Predicting that someone will duck if you throw a brick at him is easy from the folk-psychological stance; it is and will always be intractable if you have to trace the protons from brick to eyeball, the neurotransmitters from optic nerve to motor nerver, and so forth.' (Dennett, 1991), p. 42

In organisations, behaviour is assumed to be constrained by the organisational structure (e.g., (Ferber and Gutknecht, 1998; Esteva, Padget, and Sierra, 2001)),

[1] In: *Cognitive Science Quarterly* (Special Issue on Desires, Goals, Intentions, and Values: Computational Architectures), vol. 2, 2002, pp.471-494.

[2] Address for correspondence: De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Email: {jonker, treur}@cs.vu.nl, URL: http://www.cs.vu.nl/{~jonker, ~treur}

including, in particular, behavioural role specifications (cf. (Ferber, Gutknecht, Jonker, Müller, and Treur, 2001)). These role specifications enforce to a certain extent coordinated dynamics of the organisation. A role specification usually does not completely prescribe behaviours, but often allows for some space of freedom in behaviour and personal initiative. This freedom also may provide possibilities for an agent in a certain role to avoid certain behaviours as expected by others, and thus may decrease the extent of coordination. To function more efficiently in such an organisation, it is useful if agents fulfilling a certain role in the organisation can reason about the possible behaviour of the agents in other roles, for example using the intentional stance. For example, to an agent functioning within an organisation it may be very helpful to have capabilities to predict in which circumstances certain inappropriate desires or intentions are likely to arise as a basis for the behaviour of a colleague within the organisation, either

(1) to avoid the arising of these intentions by preventing the occurrence of circumstances that are likely to lead to them, or
(2) if these circumstances cannot be avoided, by anticipating consequences of the intentions.

Similarly for cases that appropriate desires or intentions may or may not arise depending on circumstances. More specific examples can be found in Section 4 below. Such capabilities of *anticipatory reasoning* about the behaviour of colleagues in an organisation are quite important for an organisation to function smoothly. This paper gives a formal basis for these types of anticipatory reasoning.

According to the intentional stance, an agent is assumed to decide to act and communicate based on its beliefs about its environment and its desires and intentions. These decisions, and the intentional notions by which they can be explained and predicted, generally depend on circumstances in the environment, and, in particular, on the information on these circumstances just acquired by observations and communication, but also on information acquired in the past. To be able to analyse the occurrence of intentional notions in the behaviour of an observed agent, the observable behavioural patterns over time form an empirical basis; cf. (Dennett, 1991).

The temporal dependencies between the intentional notions and the observable behavioural patterns, and between the intentional notions themselves, however, are only covered partially in the literature on BDI-logics as mentioned. In other references from the area of Cognitive Science and Philosophy of Mind, this omission has been criticised, and instead a different perspective is proposed, where dynamics of mental states and their interaction with the environment are central; e.g. (Bickhard, 1993, Port and van Gelder, 1995; Clark, 1997; Christensen and Hooker, 2001; Jonker and Treur, 2002). For example, (Bickhard, 1993) emphasizes 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 categorize together that class of environments that would yield that final state if interacted with. (..) the set of possible final states serves to differentiate the class of possible environments into those categories that are

> implicitly defined by the particular final states. (..) 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. However, in this literature no formalisation is proposed based on this perspective. In the formalisation introduced below, applying the general approach presented in (Jonker and Treur, 2002), the temporal aspect of the dynamics of the interaction with the environment is made explicit and related to the dynamics of beliefs, desires and intentions. Moreover, in this paper it is not assumed that the agent actually has internal states corresponding to the intentional notions. The approach makes use of the third person perspective of an external observer who attributes intentional properties to an agent. However, in Section 3.3, as a special case the relation of the presented approach to internal intentional states if these exist, is described.

Received information (observed or communicated), and decisions to perform specific actions (or communications), constitute the input and output (interface) states of an agent to the environment in which the agent functions. Externally observed behaviour traces of the agent are formalised as temporal sequences of the agent’s input and output states. The Temporal Trace Language TTL introduced in (Jonker and Treur, 1998) is used to express properties on behaviour. In this language, a (temporal) statement on the past in terms of the agent's input and output states defines a class of (possible) interaction histories; cf. (Jonker and Treur, 2002). Formal criteria are identified that express when a (temporal) statement on the past defines a class of interaction histories that can be related to a specific belief, desire or intention. A temporal statement satisfying these criteria for a specific intentional notion is viewed as a (*historical*) *temporal representation* or *temporal grounding* of this notion. These criteria can be used to identify a past statement on the agent’s interaction that can serve as an external representation. Given observations of interaction histories, the actual search for such a temporal statement can be a time-consuming computational process involving the inspection of a large number of such histories (comparable to a specific type of machine learning). However, once such an external representation statement has been identified, it can be stored and applied again and again in new situations in a very efficient manner, just by checking the current trace (or some possible trace variants, in case some impact is desirable on the occurrence of the actual trace) against this (given) statement. The approach has been implemented in an agent architecture that is capable of automatically identifying beliefs, desires and intentions of an(other) agent based on observed behaviour.

In Section 2 the language TTL used in this paper is briefly described. In Section 3, the assumptions made on the notions belief, desire and intention, and the way they interact with each other and with external notions are discussed and formalised: formal relationships between the intentional notions, and the external behaviour of an agent are defined. Formal criteria are presented that must be satisfied by a candidate temporal statement to be a justified grounding of a specific intentional notion. In Section 4 an example application in the context of organisation modelling is addressed. Section 5 is a discussion.

## 2 Basic Concepts Used

A basic assumption on the ontologies that describe properties of states of the world is that for each agent that is distinguished within the world, specific (sub)sets of ontologies of basic (atomic) world properties can be identified, according to properties that concern

- world state aspects internal to the agent,
- world state aspects external to the agent, or
- interaction aspects (input or output states of the agent).

On the basis of this assumption, ontologies for the agent's input, output and internal state are used, and for the state of the world external to the agent. It is assumed that state properties based on these ontologies describe the world state.

In the formalisation, for simplicity, we use predicate logic to specify both ontologies and properties. An ontology is specified as a finite set of sorts, constants (names) within these sorts, and relations and functions over these sorts (sometimes also called a signature). The union of two ontologies is again an ontology. For a given state ontology, state properties are the (ground) propositions that can be expressed using the concepts of an ontology. A state property is called atomic if no propositional connectives (i.e., *and*, *or*, *implies*, *not*) are used to express it.

The text below can be read without involving the formal details. To this end the formal details have been put aside in boxes, to be read only by readers interested in all technical details. For more conceptually interested readers, the text without the boxes should be readable as an independent conceptual text.

### 2.1 State Language

First, a language is used to represent facts concerning the actual state of the external world: ontology EWOnt. Some of the other (agent) ontologies will make use of EWOnt. Next, a language is used to represent facts concerning the state of the agent. The *agent input ontology* InOnt contains concepts for observation results and communication received. The following *input properties* are used for a given agent:

| | |
|---|---|
| - a property expressing the observation result that some world statement holds; e.g., it rains | denoted by observation_result(p)<br>where p denotes a state property of the external environment based on the ontology EWOnt |
| - a property expressing that agent C has communicated some world statement; e.g., agent C says to me that it rains | denoted by communicated_by(p, C)<br>where p denotes a state property of the external environment based on the ontology EWOnt |

Similarly, the *agent output ontology* OutOnt contains concepts to represent decisions to do actions within the external world, as well as concepts for outgoing communication and observations that the agent needs to obtain. The following *output properties* are used:

| | |
|---|---|
| - a property expressing that the agent decides to perform action A; e.g., take an umbrella | denoted by to_be_performed(A) |
| - a property expressing that the agent communicates information to an agent C; e.g., I say to agent C that it rains | denoted by to_be_communicated_to(p, C) where p denotes a state property of the external environment based on the ontology EWOnt |
| - a property expressing that the agent decides to perform an observation to investigate the truth of a world state property; e.g., check whether it rains | denoted by by to_be_observed(p) where p denotes a state property of the external environment based on the ontology EWOnt |

All state properties introduced to model the interaction of the agent with its environment are meta-properties: some of their arguments refer to state properties in an object-level language based on the ontology EWOnt. The *internal agent ontology* IntOnt is used for the internal (e.g., BDI) notions. The *agent interface ontology* is defined by InterfaceOnt = InOnt ∪ OutOnt; the *agent ontology* by AgOnt = InOnt ∪ IntOnt ∪ OutOnt, and the overall ontology by OvOnt = AgOnt ∪ EWOnt. In this paper we do not assume internal intentional state properties; therefore we will not assume anything particular on IntOnt. However, in Section 3.3 as a special case the existence of internal intentional state properties is discussed. The properties based on the overall state ontology are called *state properties*. All state properties based on a certain ontology Ont constitute the set SPROP(Ont).

### 2.2 Temporal Trace Language TTL

Behaviour is described by changing states over time. It is assumed that a state is characterised by the properties that hold in the state and those that do not hold.

Therefore, a *state* for ontology Ont is defined as an assignment of truth values to the set of atomic properties for Ont. The set of all possible states for ontology Ont is denoted by IS(Ont). We assume the time frame is the set of natural numbers or a finite initial segment of the natural numbers. An *overall trace* $\mathcal{M}$ over a time frame T is a sequence of states over the overall ontology OvOnt over time frame T. A *temporal domain description* $\mathcal{W}$ is a set of overall traces. Temporal domain descriptions can be compared to the information a biologist gathers on an animal by repeatedly studying its behaviour in various circumstances.

An *overall trace* $\mathcal{M}$ over a time frame $\mathbf{T}$ is a sequence of states $(M^t)_{t \in \mathbf{T}}$ in IS(OvOnt). Given an overall trace $\mathcal{M}$, the state of the input interface of agent A at time point t is denoted by

state($\mathcal{M}$, t, input(A)).

Analogously,

state($\mathcal{M}$, t, output(A))

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

state($\mathcal{M}$, t, internal(A))

the internal state. We can also refer to the overall state of a system (agents and environment) at a certain moment; this is denoted by

state($\mathcal{M}$, t).

States can be related to state properties via the satisfaction relation that expresses which properties hold in which state (comparable to the holds-relation in situation calculus); e.g., at 5 pm it rained. The *temporal trace language* TTL is built on these satisfaction relations, using the usual logical connectives and quantification (for example, over traces, time and state properties). Quantification over these entities makes the language quite expressive. For example, it allows for comparison of different possible histories in statements such as 'exercise improves skill'.

The satisfaction relation is denoted by |=. If $\varphi \in$ SPROP(InOnt), then

state($\mathcal{M}$, t, input(A)) |= $\varphi$

denotes that $\varphi$ is true in this state at time point t, based on the strong Kleene semantics (e.g., (Blamey, 1986)). The set TS(Ont) is the set of all temporal statements that only make use of ontology Ont. We allow additional language elements as abbreviations of statements of the temporal language.

To focus on different aspects of the agent, world, and time, we need ways to restrict traces. Restrictions have two parameters, one for the state ontologies and one for the time interval. The ontology parameter indicates which parts of the agent and/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 time frame of interest. Restriction is a useful way to consider the dynamics of part of the agent or world in the context of an overall trace. It allows to consider agent and world dynamics in integration: 'putting brain, body and world together again' (Clark, 1997).

For example, the notation $\mathcal{M}_{[0, t)}^{\text{InterfaceOnt}}$ denotes the restriction of $\mathcal{M}$ to the past up to t and to external state properties. The *restriction*

$\mathcal{M}_{\text{Interval}}^{\text{Ont}}$

of a trace $\mathcal{M}$ to time in Interval and information based on Ont is defined as follows:

$\mathcal{M}_{\text{Interval}}^{\text{Ont}}(t)(a) = \mathcal{M}(t)(a)$
if $t \in$ Interval and
a is an atomic property over Ont
$\mathcal{M}_{\text{Interval}}^{\text{Ont}}(t)(a) =$ unknown
otherwise

A *past statement* for trace variable $\mathcal{M}$ and time variable t is a temporal statement $\psi(\mathcal{M}, t)$ such that each time variable different from t is restricted to the time interval before t. The set of past statements over ontology Ont w.r.t. $\mathcal{M}$ and t is denoted by PS(Ont, $\mathcal{M}$, t).

For past statements, for every time quantifier for a variable t' a restriction of the form $t' \leq t$, or $t' < t$ is required within the statement. Note that for any past statement $\psi(\mathcal{M}, t)$ it holds: $\forall \mathcal{M} \in$ W $\forall t$ $\psi(\mathcal{M}_{[0, t]}, t) \Leftrightarrow \psi(\mathcal{M}, t)$.

To express that some state property has just become true, we use the qualifier *just*, denoted by ⊕. This is definable in other temporal terms: a state property has *just* become true at time t1 if and only if it is true at t1 and for some interval before t1 it was not true. Similarly it can be expressed that a state property just stopped to be true.

$$\oplus state(\mathfrak{M}, t1, interface) \models \varphi \equiv state(\mathfrak{M}, t1, interface) \models \varphi \wedge \exists t2 < t1 \ \forall t [\, t2 \leq t < t1 \Rightarrow state(\mathfrak{M}, t, interface) \models \varphi]$$

$$\oplus state(\mathfrak{M}, t1, interface) \models \varphi \equiv state(\mathfrak{M}, t1, interface) \models \varphi \wedge \exists t2 < t1 \ \forall t [\, t2 \leq t < t1 \Rightarrow state(\mathfrak{M}, t, interface) \models \varphi$$

## 3 External Representations of Beliefs, Desires and Intentions

In this section, the assumed notions of belief, desire, and intention, and their interdependencies (see Fig. 1) are discussed and formalised. The assumptions made keep the notions relatively simple; they can be extended to more complex notions. Agents are considered to which external representations of intentional properties can be attributed from a third person perspective. But also the special case that the agent has internal intentional properties is covered; more details for this case can be found in Section 3.3. The interdepenedencies depicted in Figure 1 will be interpreted as temporal interdependencies. Statements expressed in the temporal language defined above will be analysed on whether or not they are adequate candidates to express these interdependencies of intentional notions. In particular, conditions are given that formalise when a temporal statement represents a belief, desire or intention.

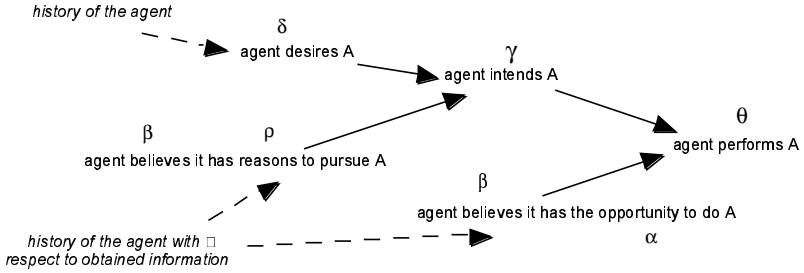


**Figure 1** Relationships between the BDI notions

A basic assumption made is that an agent's states functionally depend on the history of the agent; i.e., two copies of the same agent build up exactly the same (internal) states if they have exactly the same histories of input. For a software agent, running on a deterministic machine, this Determinism Assumption can be considered a reasonable assumption. Differences between the behaviours of two copies of the

same software agent will be created by their different histories. For most of the concepts defined below, this assumption is not strictly necessary. However, it is an assumption that strongly motivates the approach. If determinism is assumed it makes sense to exploit temporal statements that describe the history of the agent as candidates for representations of externally attributed intentional notions and actions; otherwise this approach is likely to fail.

### 3.1 Beliefs

The first intentional notion to consider is the notion of belief, which usually is considered as an informational attitude, in contrast to motivational attitudes such as desires and intentions. Viewed from the temporal perspective an agent's beliefs originate from a history of experiences; for example, observations and received communications. Also dependencies on other intentional notions via internal processes of derivation or assumption making can play a role. Moreover, beliefs affect future actions of the agent via their impact on other intentional notions (e.g., intentions or desires) that form the basis of actions. In our formalisation, in the first place beliefs are related to their history. Their relation to the future is addressed in the temporal description of the other, motivational attitudes.

In the simplest approach, beliefs ($\beta$) are based on information the agent has received by observation or communication in the past, and that has not been overridden by more recent information. This entails the first of our assumptions on beliefs: if the agent has received input in the past about a world fact, and no opposite input has been received since then, then the agent believes this world fact. The second assumption is the converse: for every belief on a world fact, there was a time at which the agent received input about this world fact (by sensing or communication), and no opposite input was received since then.

Before giving a temporal characterisation of the notion of belief an auxiliary definition is presented. The agent Ag gets information about a state property as *input* at time t if and only if it just received it at time t as an observation result or as information communicated by another agent B. This means that the agent has just received input that the state property is true at time point t.

Formally, let $p \in$ SPROP(Ont), then:

$$\texttt{Input}(p, t, \mathfrak{M}, Ag) \equiv$$
$$\oplus\text{state}(\mathfrak{M}, t, \text{input}(Ag)) \models \text{observation\_result}(p)$$
$$\vee\ \exists B \in \text{AGENT}\ \oplus\text{state}(\mathfrak{M}, t, \text{input}(Ag)) \models \text{communicated\_by}(p, B)$$

Here AGENT is a sort for the agent names. For simplicity of notation, often the fourth argument Ag will be left out:

$$\texttt{Input}(p, t, \mathfrak{M})$$

**Definition (Temporal Belief Statement)**

The following characterisation of belief is based on the assumption that an agent believes a fact if and only if it received input about it in the past and the fact is not contradicted by later input of the opposite. Let $\alpha \in$ SPROP(Ont) be a state property over Ont. The temporal statement $\beta(\mathfrak{M}, t) \in$ TS is a *temporal belief statement* for state property $\alpha$ if and only if:

at each time point t and each trace $\mathcal{M}$ the statement $\beta(\mathcal{M}, t)$ is true if and only if at an earlier time point t1 the agent received input that $\alpha$ is true and after this time point did not receive input that $\alpha$ is false. Sometimes this belief statement is denoted by $\beta_\alpha(\mathcal{M}, t)$, to indicate it is a belief statement for $\alpha$.

In the specific case that $\beta(\mathcal{M}, t)$ is a temporal belief statement, and, in addition, $\beta(\mathcal{M}, t)$ is a temporal past statement (i.e., $\beta(\mathcal{M}, t) \in PS(InOnt, \mathcal{M}, t)$), over ontology InOnt, then it is also called a *historical belief statement* for $\alpha$.

Note that one particular historical belief statement for $\alpha$ is the temporal past statement $\mathtt{Belief}(\alpha, t, \mathcal{M}) \in PS(InOnt, \mathcal{M}, t)$ stating that 'at an earlier time point the agent received input that $\alpha$ is true and after this time point did not receive input that $\alpha$ is false'.

Formally, the temporal statement $\beta(\mathcal{M}, t1) \in TS$ is a *temporal belief statement* for $\alpha$ if and only if

$\forall \mathcal{M} \in \mathcal{W}\ \forall t1\ [\beta(\mathcal{M}, t1) \Leftrightarrow$
$\exists t0 \leq t1\ [\ \mathtt{Input}(\alpha, t0, \mathcal{M}) \wedge$
$\forall t \in [t0, t1]\ \neg\ \mathtt{Input}(\sim\alpha, t, \mathcal{M})\ ]\ ]$

Here for state property $\alpha$, the *complementary property* $\sim\alpha$ is defined as

$\sim\alpha = \alpha'$ if $\alpha = \neg\alpha'$
$\sim\alpha = \neg\alpha$ otherwise

The temporal past statement
$\mathtt{Belief}(\alpha, t, \mathcal{M}) \in PS(InOnt, \mathcal{M}, t)$
is formally defined by
$\exists t0 \leq t\ [\ \mathtt{Input}(\alpha, t0, \mathcal{M})\ \wedge$
$\forall t1 \in [t0, t]\ \neg\ \mathtt{Input}(\sim\alpha, t1, \mathcal{M})\ ]$

If required, these assumptions can also be replaced by less simple ones, possibly in a domain-dependent manner; for example, taking into account reliability of sensory processes in observation or reliability of other agents in communication. As the criterion for being a temporal belief statement is exactly the criterion expressed as the definition of $\mathtt{Belief}(p, t, \mathcal{M})$, for every belief statement it holds at a point in time t precisely if $\mathtt{Belief}(p, t, \mathcal{M})$ holds at t. In this sense all belief statements are temporally equivalent. Moreover, under the assumption that input atoms are always correct with respect to the actual world state, it is not possible to receive observation or comunication that a world state property p is true and is false at the same point in time. This entails that it is not possible to have at any point in time contradictory beliefs. These results are summarized in the following proposition.

**Proposition 3.1**

Let a state property $p \in SPROP(Ont)$ be given.

a) All temporal belief statements for p are temporally equivalent; i.e., if $\beta1(\mathcal{M}, t)$ and $\beta2(\mathcal{M}, t) \in TS$ are two temporal belief statements for p, then:

for each trace and time point one is true if and only if the other is true.

Formally, *temporally equivalent* means:
$\forall \mathcal{M} \in \mathcal{W}\ \forall t \quad \beta1(\mathcal{M}, t) \Leftrightarrow \beta2(\mathcal{M}, t)$

b) Suppose the input atoms are correct with respect to the world state. Then at each time point there are no belief

An input atom is called *correct* with respect to the world state if and only if
$\forall \mathcal{M} \in \mathcal{W}\ \forall t\ \forall p$

statements true for complementary world state properties. In other words, for any world state property p, if $\beta_1(\mathcal{M}, t)$ is a temporal belief statement for p and $\beta_2(\mathcal{M}, t)$ is a temporal belief statement for the complementary property ~ p, then these two belief statements exclude each other, i.e., if in a given trace and at a given time point one of the temporal belief representations is true, then the other is false.

$[\ \mathtt{Input}(p, t, \mathcal{M}) \Rightarrow \mathrm{state}(\mathcal{M}, t, EW) \models p\ ]$

Formally, $\beta_1(\mathcal{M}, t)$ and $\beta_2(\mathcal{M}, t)$ *exclude each other* means

$\forall \mathcal{M} \in \mathfrak{W}\ \forall t\ \ \beta_1(\mathcal{M}, t) \ \Rightarrow\ \neg\, \beta_2(\mathcal{M}, t)$

### 3.2 Desires and Intentions

Also motivational attitudes can be viewed from a temporal perspective. Although they also have past, motivational attitudes refer in their semantics in a generic manner to the future actions of the agent, so it can be expected that in a temporal characterisation a reference to future actions of the agent is made. Our assumptions on intentions are as follows. In the first place, under appropriate circumstances an intention leads to an action: an agent who intends to perform an action will execute the action when an opportunity ($\alpha$) occurs. Moreover, the second assumption is that when an action or communication (A) is performed ($\theta$), the agent is assumed to have intended ($\gamma$) to do that.

**Definition (Temporal Intention Statement)**

An *action atom* $\theta(\mathcal{M}, t, Ag)$ is an atom stating that at time point t in trace $\mathcal{M}$ at the output of the agent Ag a specific generated action or communication can be found.

Formally, an *action atom* $\theta(\mathcal{M}, t, Ag)$ is an atom of the form

$\mathrm{state}(\mathcal{M}, t, \mathrm{output}(Ag)) \models \psi$

with $\psi$ an output atom: an atom of the form to_be_performed(A), to_be_communicated_to(p, B), or to_be_observed(p).

Let $\alpha \in$ SPROP(EWOnt) be an external state property and $\theta(\mathcal{M}, t, Ag)$ an action atom. The temporal statement $\gamma(\mathcal{M}, t) \in$ TS is called a *temporal intention statement* for action atom $\theta(\mathcal{M}, t, Ag)$ and *opportunity* $\alpha$ if and only if the following conditions are fulfilled:

*Sufficiency condition for intention*

If $\gamma(\mathcal{M}, t)$ holds for a given trace $\mathcal{M}$ and time point t1, and at some earlier time point the agent received input that $\alpha$ holds and since then the agent did not receive input that $\alpha$ does not hold, then there is a time point t2 later than t1 at which the action $\theta(\mathcal{M}, t2, Ag)$ occurs.

Formally, the *sufficiency condition for intention* is defined by:

$\forall \mathcal{M} \in \mathfrak{W}\ \forall t1\ [\ \gamma(\mathcal{M}, t1)\ \wedge$
$\exists t0 \le t1\ [\ \mathtt{Input}(\alpha, t0, \mathcal{M})\ \wedge$
$\forall t \in [t0, t1]\ \neg\, \mathtt{Input}(\sim\alpha, t, \mathcal{M})\ ]$
$\Rightarrow\ \exists t2 \ge t1\ \theta(\mathcal{M}, t2, Ag)\ ]$

*Necessity condition for intention:*

If for a given trace $\mathcal{M}$ and time point t2 the action $\theta(\mathcal{M}, t, Ag)$ occurs, then $\gamma(\mathcal{M}, t1)$ holds at some earlier time point t1 and at a time point earlier than t1 the agent received input that $\alpha$ holds and since then until t1 the agent did not receive input that $\alpha$ does not hold.

Formally, the *necessity condition for intention* is defined by:

$$\forall \mathcal{M} \in \mathcal{W}\ \forall t2\ [\ \theta(\mathcal{M}, t2, Ag) \Rightarrow \exists t1 \leq t2\ \gamma(\mathcal{M}, t1)\ \wedge \exists t0 \leq t1\ [\ \mathtt{Input}(\alpha, t0, \mathcal{M}) \wedge \forall t \in [t0, t1]\ \neg\, \mathtt{Input}(\sim\alpha, t, \mathcal{M})]]$$

In the specific case that the past statement $\gamma_P(\mathcal{M}, t) \in PS(InOnt, \mathcal{M}, t)$ is a temporal intention statement for $\theta(\mathcal{M}, t, Ag)$ and opportunity $\alpha$, it is also called a *historical intention statement* for action atom $\theta(\mathcal{M}, t, Ag)$ and opportunity $\alpha$.

The above definition formalises the case that all actions are intended actions. However, it is not difficult to define weaker variants. For example, if also unintended actions are allowed, the second (necessity) condition can be left out.

If for the external state property $\alpha$ used for the opportunity, any temporal belief statement $\beta_\alpha(\mathcal{M}, t)$ is given, then the characterisation of an intention can be reformulated by replacing the clause

'at some earlier time point the agent received input that $\alpha$ holds and since then the agent did not receive input that $\alpha$ does not hold'

by $\beta_\alpha(\mathcal{M}, t)$. This allows simplification, as is summarized in the following proposition.

**Proposition 3.2**

Let $\alpha \in SPROP(EWOnt)$ be an external state property, $\beta_\alpha(\mathcal{M}, t)$ be a belief statement for $\alpha$ and $\theta(\mathcal{M}, t, Ag)$ an action atom. The temporal statement $\gamma(\mathcal{M}, t) \in TS$ is a temporal intention statement for action atom $\theta(\mathcal{M}, t, Ag)$ and opportunity $\alpha$ if and only if the following conditions are fulfilled:

*Sufficiency condition for intention:*

If $\gamma(\mathcal{M}, t1)$ and $\beta_\alpha(\mathcal{M}, t1)$ both hold for a given trace $\mathcal{M}$ and time point t1, then there is a time point t2 later than t1 at which the action $\theta(\mathcal{M}, t2, Ag)$ occurs.

Formally, the sufficiency condition for intention is reformulated into:

$$\forall \mathcal{M} \in \mathcal{W}\ \forall t1\ [\gamma(\mathcal{M}, t1)\ \wedge\ \beta_\alpha(\mathcal{M}, t1) \Rightarrow\ \exists t2 \geq t1\ \theta(\mathcal{M}, t2, Ag)\ ]$$

*Necessity condition for intention:*

If for a given trace $\mathcal{M}$ and time point t2 the action $\theta(\mathcal{M}, t2, Ag)$ occurs, then an earlier time point t1 exists for which both $\gamma(\mathcal{M}, t1)$ and $\beta_\alpha(\mathcal{M}, t1)$ hold.

The necessity condition for intention is reformulated into:

$$\forall \mathcal{M} \in \mathcal{W}\ \forall t2\ [\ \theta(\mathcal{M}, t2, Ag) \Rightarrow\ \exists t1 \leq t2\ \gamma(\mathcal{M}, t1)\ \wedge \beta_\alpha(\mathcal{M}, t1)\ ]$$

The following simple example illustrates the notions introduced. The observed (animal) agent receives observation input on the availability of food (food), and of the

limitation of its moving around due to the presence or absence of a screen in a certain experimental setting (screen). Depending on the circumstances it can decide to eat the food (action eat). Assume that the traces depicted in Table 1 are observed.

| *time*<br>*trace* | *time point*<br>*0* | *time point*<br>*1* | *time point*<br>*2* | *time point*<br>*3* | *time point*<br>*4* | *time point*<br>*5* |
|---|---|---|---|---|---|---|
| *trace 1* | food<br>screen | no food<br>no screen | food<br>screen | food<br>no screen | food<br>no screen<br>eat | food<br>no screen<br>eat |
| *trace 2* | no food<br>no screen | food<br>no screen | no food<br>screen | food<br>no screen | food<br>no screen | food<br>no screen<br>eat |
| *trace 3* | no food<br>no screen | no food<br>no screen | food<br>screen | food<br>no screen | food<br>no screen | food<br>no screen |
| *trace 4* | food<br>no screen | no food<br>no screen | food<br>screen | food<br>screen | food<br>no screen | food<br>no screen<br>eat |

**Table 1** Example set of observed traces

For the state property no screen as opportunity, the following past statement $\gamma_P(\mathcal{M}, t)$ was found to be an adequate intention representation:

at t the agent observes the presence of food, and there exist time points t1 before t and t2 before t1 such that the agent observed the absence of food at t1 and the presence of food at t2. Informally this statement can be explained as follows: the agent has the intention to eat at each time point that food is visible and in the past the agent experienced that visible food can suddenly disappear.

Formally:

$\gamma_P(\mathcal{M}, t) =$

$state(\mathcal{M}, t, input(agent)) \models$
$\quad observation\_result(food)$
$\wedge\ \exists t1 \leq t\ [\ state(s, \mathcal{M}, t1, input(agent)) \models$
$\quad observation\_result(\neg\ food)$
$\quad \wedge\ \exists t2 \leq t1\ state(s, \mathcal{M}, t2, input(agent)) \models$
$\quad observation\_result(food)\ ]$

An agent can desire states of the world as well as actions to be performed. When the agent has a set of desires, it can choose to pursue some of them. A chosen desire for a state of the world can lead to an intention to do an action if, for example, expected effects of the action (partly) fulfil the desire. The first assumption on desires is that, given a desire ($\delta$), for each relevant action there is an additional reason ($\rho$), so that if both the desire is present and the agent believes the additional reason, then the intention to perform the action will be generated. Having this additional reason prevents the agent from performing actions that do not make sense in the given situation; e.g., actions with contradicting effects. The second assumption formalised in the definition below is that every intention is based on a desire ($\delta$), i.e., no intention occurs without desire. Based on these assumptions, temporal desire statements are defined as follows:

**Definition (Temporal Desire Statement)**
Let an external state property $\rho \in$ SPROP(EWOnt) and an intention statement $\gamma(\mathcal{M}, t)$ be given. The temporal statement $\delta(\mathcal{M}, t) \in$ TS is called a *temporal desire statement* for intention $\gamma(\mathcal{M}, t)$ and *additional reason* $\rho$ if and only if the following conditions are fulfilled:

*Sufficiency condition for desire*

If $\delta(\mathcal{M}, t1)$ holds for a given trace $\mathcal{M}$ and time point t1, and at some earlier time point the agent received input that $\rho$ holds and since then the agent did not receive input that $\rho$ does not hold, then there is a time point t2 later than t1 at which the intention $\gamma(\mathcal{M}, t2)$ occurs.

Formally, the *sufficiency condition for desire* is defined by:

$\forall \mathcal{M} \in \mathcal{W}\ \forall t1\ [\delta(\mathcal{M}, t1) \wedge$
$\exists t0 \leq t1\ [\mathtt{Input}(\rho, t0, \mathcal{M}) \wedge$
$\forall t \in [t0, t1]\ \neg\ \mathtt{Input}(\sim\rho, t, \mathcal{M})\ ]$
$\Rightarrow\ \exists t2 \geq t1\ \gamma(\mathcal{M}, t2)\ ]$

*Necessity condition for desire:*

If for a given trace $\mathcal{M}$ and time point t2 the intention $\gamma(\mathcal{M}, t2)$ occurs, then the desire $\delta(\mathcal{M}, t1)$ holds at some earlier time point t1 and at a time point earlier than t1 the agent received input that $\rho$ holds and since then until t1 the agent did not receive input that $\rho$ does not hold.

Formally, the *necessity condition for desire* is defined by:

$\forall \mathcal{M} \in \mathcal{W}\ \forall t2\ [\gamma(\mathcal{M}, t2) \Rightarrow$

$\exists t1 \leq t2\ \ \delta(\mathcal{M}, t1) \wedge \exists t0 \leq t1\ [\ \mathtt{Input}(\rho, t0, \mathcal{M}) \wedge \forall t \in [t0, t1]\ \neg\ \mathtt{Input}(\sim\rho, t, \mathcal{M})\ ]\ ]$

If the past statement $\delta_P(\mathcal{M}, t) \in$ PS(InOnt, $\mathcal{M}$, t) is a temporal desire statement for intention $\gamma(\mathcal{M}, t)$ and additional reason $\rho$, it is called a *historical desire statement* for intention $\gamma(\mathcal{M}, t)$ and (additional) *reason* $\rho$.

As for intentions, weaker notions can be defined as well. For example, the second assumption, that no intentions occur without desire, may be debatable. If also undesired intentions are allowed, this assumption can be dropped by leaving out the second (necessity) condition of the above definition.

If for the external state property $\rho$ used, any temporal belief statement $\beta_\rho(\mathcal{M}, t)$ is given, then the characterisation of a desire can be reformulated by replacing the clause

> 'at some earlier time point the agent received input that $\rho$ holds and since then the agent did not receive input that $\rho$ does not hold'

by $\beta_\rho(\mathcal{M}, t)$. This allows simplification, as is summarized in the following proposition.

**Proposition 3.3**

Let $\rho \in$ SPROP(EWOnt) be an external state property, $\beta_\rho(\mathcal{M}, t)$ a belief statement for $\rho$ and $\gamma(\mathcal{M}, t)$ an intention statement. The temporal statement $\delta(\mathcal{M}, t) \in$ TS is a temporal desire statement for intention $\gamma(\mathcal{M}, t)$ and additional reason $\rho$ if and only if the following conditions are fulfilled:

*Sufficiency condition for desire:*

If $\delta(\mathcal{M}, t1)$ and $\beta_p(\mathcal{M}, t1)$ both hold for a given trace $\mathcal{M}$ and time point t1, then there is a time point t2 later than t1 at which the intention $\gamma(\mathcal{M}, t2)$ occurs.

*Necessity condition for desire:*

If for a given trace $\mathcal{M}$ and time point t2 the intention $\gamma(\mathcal{M}, t2)$ occurs, then an earlier time point t1 exists for which both $\delta(\mathcal{M}, t1)$ and $\beta_p(\mathcal{M}, t1)$ hold.

Formally, the sufficiency condition for desire is reformulated into:

$$\forall \mathcal{M} \in \mathcal{W}\ \forall t1\ [\delta(\mathcal{M}, t1)\ \wedge \beta_p(\mathcal{M}, t1) \Rightarrow \exists t2 \geq t1\ \gamma(\mathcal{M}, t2)\ ]$$

The necessity condition for desire is reformulated into:

$$\forall \mathcal{M} \in \mathcal{W}\ \forall t2\ [\gamma(\mathcal{M}, t2) \Rightarrow \exists t1 \leq t2\ \ \delta(\mathcal{M}, t1)\ \wedge \beta_p(\mathcal{M}, t1)\ ]$$

Returning to the animal behaviour example, for the state property food as additional reason, the following past statement $\delta_P(\mathcal{M}, t)$ was found to be an adequate temporal desire statement for the example intention statement specified above:

at t the agent observes the absence of food, and there exist a time point t1 before t such that the agent observed the presence of food at t1. Informally this statement can be explained as follows:

Formally: $\delta_P(\mathcal{M}, t) =$

$$\text{state}(\mathcal{M}, t, \text{input(agent)}) \models \text{observation\_result}(\neg\text{food})$$
$$\wedge\ \exists t1 \leq t\ [\ \text{state}(s, \mathcal{M}, t1, \text{input(agent)}) \models \text{observation\_result}(\text{food})$$

the agent has the desire to eat at each time point that the absence of food is observed and in the past the agent observed the presence of food. In other words, the agent desires what is no longer present.

### 3.3 Internal Representations

In this paper no internal (mental) concepts of an agent are assumed. However, if internal notions of belief, desire and intention of an agent happen to exist, or at least are claimed (e.g., because the agent was designed and implemented this way), our framework can be applied to them as well. If internal representations of beliefs, desires and intentions of an agent happen to exist, during execution these internal representations change according to interaction and internal processing of the agent. In this case the temporal relationships between an internal representation and properties of the history and future of the agent's interaction (obtained information and performed actions), that is, the external behaviour of the agent can be formulated on the basis of the criteria introduced.

**Definition (Internal Representations)**

a) The internal state property $\beta \in$ SPROP(IntOnt) is called an *internal belief representation* for state property p if the temporal statement expressed by:

within trace $\mathcal{M}$ at time point t state property $\beta$ holds in the agent's internal state

is a temporal belief statement for p.

Formally this is expressed as:

$$\text{state}(\mathcal{M}, t, \text{internal(Ag)}) \models \beta$$

b) The internal state property $\gamma \in$ SPROP(IntOnt) is called an *internal intention representation* for action statement $\theta(\mathcal{M}, t, Ag)$ and opportunity $\alpha$ if and only if the statement

within trace $\mathcal{M}$ at time point t state property $\gamma$ holds in the agent's internal state

Formally:
$state(\mathcal{M}, t, internal(Ag)) \models \gamma$

is a temporal intention statement for $\theta(\mathcal{M}, t)$ and opportunity $\alpha$.

c) The internal state property $\delta \in$ SPROP(IntOnt) is called an *internal desire representation* for intention $\gamma(\mathcal{M}, t)$ and *reason* $\rho$ if and only if the temporal statement

within trace $\mathcal{M}$ at time point t state property $\delta$ holds in the agent's internal state

Formally:
$state(\mathcal{M}, t, internal(Ag)) \models \delta$

is a temporal desire statement for intention $\gamma(\mathcal{M}, t)$ and reason $\rho$.

## 4 Anticipatory Reasoning and Acting in Organisations

Viewed from a dynamic perspective, organisational structure (cf. (Ferber and Gutknecht, 1998; Esteva, Padget, and Sierra, 2001)), provides specifications of constraints on the dynamics of role behaviour and interactions (cf. (Ferber et al., 2001)). By these specifications to a certain extent coordinated dynamics is enforced to the organisation. In human organisations role specifications usually do not completely prescribe behaviours, however. To a greater or lesser extent some space of freedom in behaviour and personal initiative is allowed. This freedom has its positive elements; in the first place, human agents can find more satisfaction and do their work with higher quality if they can do things in their own way. In the second place an organisational structure does not anticipate on all possible circumstances; in unforeseen situations it can be beneficial if agents have some space to improvise.

The obverse, however, is that this freedom also may provide possibilities to agents to show certain behaviours (based on their individual characteristics and interests) that decrease the extent of coordination. To function more efficient in an organisation, where roles do not completely prescribe behaviour, it is useful if agents fulfilling a certain role in the organisation can reason in an anticipatory sense about the behaviour of the agents in other roles, for example, using the intentional stance. This section addresses this application of the framework introduced in Section 3 in more detail. Some examples of the phenomena described for human organisations are:

(a) An employee has done something very important very wrong, and deliberates whether or not to tell his manager: *'If he believes that I am the cause of the problems, he will try to fire me.'*

(b) An employee has encountered a recurring problem, and knows a solution for this problem, on which he would like to work. He deliberates about how to propose to his manager this solution. *'If I tell this solution immediately he will not believe that the problem is worth working on it. If I make him aware of the problem, and do not tell a solution, he only will start to think himself about it for a while, without finding a solution,*

*and then forget about it. If I make him aware of the problem and give some hints that direct him to a (my) solution, he will believe he contributed to a solution himself and want me to work on it.'*

(c) A manager observes that a specific employee in the majority of cases functions quite cooperatively, but shows avoidance behaviour in other cases. In these latter cases, the employee starts trying to reject the task if he believes that his agenda already was full-booked for the short term, it is not clear to him whether somebody else is not capable of doing the task, and he believes colleagues are available with less full-booked agendas. Further observation by the manager reveals the pattern that the employee shows avoidance behaviour, in particular, in cases that a task is only asked shortly before its deadline, without the possibility to anticipate on the possibility of having the task allocated. The manager deliberates about this as follows: *'If I know beforehand the possibility that a last-minute task will occur, I can tell him the possibility in advance, and in addition point out that I need his unique expertise for the task, in order to avoid the behaviour that he tries to avoid the task when it actually comes up.'*

The reasoning processes on predicted behaviours described in (a) to (c) can be based on prescribed role behaviours (as may be the case in (a)), or on an analysis of the other agent's personal motivations (as is the case in (b) and (c)). Especially in these latter cases, the analysis framework developed in this paper is applicable. To show this, example (c) is addressed by making the following interpretation.

The *desire* to avoid a task is created after time t by the employee if the following holds for the history:
- at time t the employee heard the request to perform the task
- at time t the employee observes that the task has to be finished soon
- the employee did not hear of the possibility of the task at any earlier time point

The *intention* to avoid a task is generated after time t if the following holds for the history:
- the desire to avoid the task is available at time t
- the belief that colleagues are capable of doing the task is available at time t
- the belief that colleagues are not full-booked is available at time t

The *action* to avoid the task is generated at time t if the following holds for the history:
- the intention to avoid the task is available at time t
- the belief that the employee's own agenda is full-booked is available at time t

The formalisations of these conditions are as follows.

The *input ontology* InOnt includes:

observation_result(task_urgent),
observation_result(own_agenda_full),
observation_result(colleagues_agenda_not_full),
observation_result(colleagues_capable_of_task),
communicated(task_request),
communicated(task_possibility)

The *output ontology* OutOnt includes tbc(task_rejection). Here tbc is short for 'to be communicated'.

Define the past statement $\delta_P(\mathcal{M}, t) \in$ PS(Ont, $\mathcal{M}$, t ) for the *desire* to avoid the task by

state($\mathcal{M}$, t, input(Ag)) |= communicated(task_request) $\wedge$
state($\mathcal{M}$, t, input(Ag)) |= observation_result(task_urgent) $\wedge$
$\neg \exists t0 < t$ state($\mathcal{M}$, t0, input(Ag)) |= communicated(task_possibility)

The *additional reason* $\rho$ to generate an avoidance intention is:

colleagues_agenda_not_full $\wedge$ colleagues_capable_of_task

The past statement $\gamma_P(\mathcal{M}, t) \in$ PS(Ont, $\mathcal{M}$, t) for the *intention* to avoid the task is defined in short form by

state($\mathcal{M}$, t, input(Ag)) |= communicated(task_request) $\wedge$
state($\mathcal{M}$, t, input(Ag)) |= observation_result(task_urgent) $\wedge$
$\neg \exists t0 < t$ state($\mathcal{M}$, t0, input(Ag)) |= communicated(task_possibility) $\wedge$
$\exists t0 \leq t1$ [ Input(colleagues_agenda_not_full $\wedge$ colleagues_capable_of_task, t0, $\mathcal{M}$) $\wedge$
$\forall t \in [t0, t1]$ $\neg$ Input(~(colleagues_agenda_not_full $\wedge$ colleagues_capable_of_task), t, $\mathcal{M}$)

The short form is

$\delta_P(\mathcal{M}, t)$ $\wedge$ Belief(colleagues_agenda_not_full $\wedge$ colleagues_capable_of_task, t, $\mathcal{M}$)

The *opportunity* $\alpha$ to perform the avoidance action is:

own_agenda_full

The past statement $\theta_P(\mathcal{M}, t, Ag) \in$ PS(Ont, $\mathcal{M}$, t ) for the *action* to avoid the task is defined in its short form by

$\gamma_P(\mathcal{M}, t)$ $\wedge$ Belief(own_agenda_full, t, $\mathcal{M}$)

Given this interpretation it can be illustrated how the manager agent can reason and act in an anticipatory manner to avoid the employee's avoidance desire, intention and/or action to occur. This can be done in the following three manners:

(1) *Avoiding the desire to occur*

This can be obtained by communicating in advance to the employee that possibly a last minute task will occur. This would make the third condition in the definition of the temporal desire statement fail.

(2) *Avoiding the intention to occur (given that the desire occurs)*

This can be obtained by refutation of the reason to generate the intention, e.g., by telling the employee that he is the only one with the required expertise.

(3) *Avoiding the action to occur (given that the intention occurs)*

This can be obtained by refutation of the opportunity, e.g., by taking one of the (perhaps less interesting) tasks from his agenda and re-allocating it to a colleague.

## 5 Discussion

In this section a number of more specific themes are discussed in different subsections.

### 5.1 On fundamental assumptions behind the approach

Two fundamental assumptions are discussed. First, the approach presented is based on the fundamental assumption that it is known from an external observer's perspective what the agent is and what its interfaces to the external world are, in particular, what its actions, observations and communication are. This is a severe assumption, since in nontrivial cases it may be not simple at all to interpret whether, for example, some observed pattern is an agent's action or not. This problem is not addressed here. For cases that this problem can be solved, the approach put forward is applicable.

A second assumption relates to notions of causality. This paper contributes a formal analysis of the dynamics of mental properties such as beliefs, desires and intentions in the context of the dynamics of the interaction with the agent's environment. The analysis results in a number of temporal relationships between such intentional states and interaction traces. It could be asked whether these temporal relationships are meant as a specific type of causality. This is not the case. In Philosophy an appropriate notion of causality is the subject of a serious debate; e.g., (Sosa and Tooly, 1993), without a satisfactory outcome. Therefore it was decided to take a modest perspective and base our analysis on mathematically defined temporal relationships without any further claim that these relationships could count as some form of causality. In fact, the notion of temporal dependence as exploited may be positioned close to the notion based on either sufficient or necessary conditions (or both) discussed in (Sosa and Tooley, 1993), pp. 5-8:

I. C is a cause of E if and only if C and E are actual and C is ceteris paribus sufficient for E

II. C is a cause of E if and only if C and E are actual and C is ceteris paribus necessary for E

However, as the discussion in the reference mentioned clearly shows, these notions does not fulfill requirements typically inposed on a notion of causality. In fact our notions based on temporal dependence takes into account the dynamic aspects in more detail and fits more closely to the interactivist perspective (cf. (Bickhard, 1993)); see (Jonker and Treur, 2002) for more details. In summary, we do not claim our type of relationship to be a notion of causality.

### 5.2 Expressivity of the Temporal Trace Language

The temporal trace language TTL used in our approach is much more expressive than standard temporal logics in a number of respects. Therefore it can be used for different types of applications. In the first place, it has *order-sorted predicate logic* expressivity, whereas most standard temporal logics are propositional. Secondly, the explicit reference to *time points and time durations* offers the possibility of modelling the dynamics of real-time phenomena, such as sensory and neural activity patterns in relation to mental properties (cf. (Port and van Gelder, 1995)).

Third, the possibility to quantify over traces allows for specification of *more complex behaviours*. As within most temporal logics, reactiveness and pro-activeness

properties can be specified. In addition, in our language also properties expressing different types of adaptive behaviour can be expressed. For example a relative adaptive 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. This type of property can be expressed in TTL, whereas in standard forms of temporal logic, or in approaches as presented in (Ismail and Shapiro, 2000; Purang, Purushothaman, Traum, Andersen, and Perlis, 1999) different alternative histories cannot be compared. As another example, the *monotonicity property of trust* as identified and mathematically formalised in (Jonker and Treur, 1999), which roughly spoken states that 'the more positive the experiences, the higher the trust', cannot be expressed in a standard temporal logic but is expressible in TTL; cf. (Marx and Treur, 2001).

In the current paper only part of the features of the language TTL as discussed above are exploited; it is not claimed that TTL is unique for this context. Due to the simplifying assumptions on the temporal relationships between intentional notions addressed here, for this focus the job could also be done by many less expressive languages. However, then the approach is less generic and will not be extendable to more complex behaviours and mental properties.

### 5.3 Relation to BDI-models

The formal analysis presented in this paper differs from the approaches to BDI-models in e.g., (Cohen and Levesque, 1990; Linder et al., 1996; Rao and Georgeff, 1991) in that it relates in a dynamic manner intrinsically internal notions to external notions, like observations, communications and actions. Criteria for the notions belief, desire, and intention in terms of external notions are presented. The criteria allow for (1) externally ascribing motivational attitudes to agents (that may not use any belief, desire or intention internally) by defining these notions in terms of the external behaviour of the agent, and, (2) for analysis of internal notions, and (3) anticipatory reasoning to affect the circumstances that may lead to the generation of beliefs, desires and/or intentions.

### 5.4 Relation to Other Approaches

An approach that in some aspects is similar in perspective to ours, is that of (Rosenschein and Kaelbling, 1986). They ascribe knowledge to so-called situated automata, which are processes that do not have any internal representation of knowledge. A process with a certain internal state $v$ knows $\varphi$ if $\varphi$ is true in all environment situations which are possible when the process is in state $v$. Our approach for ascribing beliefs is different; we relate belief to the acquired information on the environment. Furthermore, Rosenschein and Kaelbling give no account of desire and intention, which is a main contribution of our paper. The same holds for recent work presented in (Wooldridge and Lomuscio, 2000), which concentrates on the informational aspects, and abstracts from motivational and temporal aspects; actually, in (Wooldridge and Lomuscio, 2000) exploration of the temporal aspects, as presented above, is mentioned as one of the four items on the list of issues for future work.

In research on plan recognition, such as (Allen, 1983; Konolige and Pollack, 1989; Goldman, Geib, and Miller, 1999), based on observed actions of an actor agent the observing agent ascribes intentions and plans to the actor that are probable. Plan

recognition is performed using data on the actions from a single, ongoing interaction of the agent, and uses domain knowledge on actions and their expected effects in a crucial manner. Our approach is quite different. The analysing agent primarily takes circumstances that may lead to certain intentions into account using information on the observations in the past of the actor studied, in order to find hypothetical past statements representing the beliefs, desires and intentions of this agent. No domain knowledge on actions and effects is used.

### 5.5 Further perspectives

The approach introduced here opens up a number of possibilities for further work. In the first place, the model for beliefs, desires and intentions and their dynamics can be made more complex. In particular, questions concerning revision and update of beliefs, desires and intentions can be addressed from the temporal perspective, for example taking the detailed analysis of (Bratman, 1987, 1999) as a point of departure.

It would be interesting also to explore whether variants of the formalisation of intention attribution as introduced in this paper can be related to empirical results as reported in the Cognitive Science literature. The overview of an impressive amount of empirical literature on intention attribution, contributed in (Baldwin and Baird, 2001) can be a good starting point for this future work.

## References


Allen, J.F. (1983). Recognizing intentions from natural language utterances. In M. Brady and R.C. Berwick, eds., *Computational Models of Discourse*. MIT Press, Cambridge, Ma., 1983.

Baldwin, D.A., and Baird, J.A. (2001). Discerning intentions in dynamic human action. In: *Trends in Cognitive Sciences*, vol. 5 (2001), pp. 171-178.

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

Blamey, S. (1986). Partial Logic, in: D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, Vol. III, 1-70, Reidel, Dordrecht, 1986.

Bratman, M.E., (1987). *Intention, Plans and Practical Reason.* Harvard University Press

Bratman, M.E., (1999). *Faces of Intention.* Cambridge University Press

Christensen, W.D. and Hooker, C.A. (2001). Representation and the Meaning of Life. In: *Representation in Mind: New Approaches to Mental Representation*, Proceedings, 27-29th June 2000 University of Sydney. To be published by Springer Verlag, 2002.

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

Cohen, P.R. and Levesque, H.J. (1990). Intention is Choice with Commitment. *Artificial Intelligence* vol. 42 (1990), pp. 213-261.

Dennett, D.C. (1987). *The Intentional Stance.* MIT Press. Cambridge Mass, 1987.

Dennett, D.C. (1991). Real Patterns. *The Journal of Philosophy*, vol. 88, 1991, pp. 27-51.

Esteva, M., Padget, J., and Sierra, C, (2001). Formalizing a language for institutions and norms. In: *Intelligent Agents VIII, Proc. of the 8th International Workshop on Agent Theories, and Languages, ATAL'01*. Ledcture Notres in AI, Springer Verlag, pp. 106-119.

Ferber, J. and Gutknecht, O. (1998). A meta-model for the analysis and design of organizations in multi-agent systems. *Third International Conference on Multi-Agent Systems (ICMAS '98) Proceedings.* IEEE Computer Society, 1998, pp. 128-135.

Ferber, J., Gutknecht, O., Jonker, C.M., Mueller, J.P., and Treur, J. (2001). Organization Models and Behavioural Requirements Specification for Multi-Agent Systems. In: Y. Demazeau, and F.J. Garijo (eds.), *Multi-Agent Organisations, Proceedings of the 10th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'01*, 2001 . To be published in Lecture Notes in AI, Springer Verlag, Berlin.

Fisher, M. (1994). A survey of Concurrent METATEM — the language and its applications. In: D.M. Gabbay, H.J. Ohlbach (eds.), *Temporal Logic — Proceedings of the First International Conference*, Lecture Notes in AI, vol. 827, 1994, pp. 480–505.

Goldman, R.P., Geib, C.W., and Miller, C.A. (1999). A New Model of Plan Recognition. In: *Proc. of the Conference on Uncertainty in AI*. Stockholm.

Ismail, H.O., and Shapiro, S.C. (2000). Two problems with Reasoning and Acting in Time. In: A.G. Cohn, F. Giunchiglia and B. Selman (eds.), *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning, KR 2000*. Morgan Kaufman.

Jonker, C.M., and Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: F.J. Garijo, M. Boman (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, Springer Verlag, Berlin, 1999, pp. 221-232.

Jonker, C.M., and Treur, J. (2002). A Dynamic Perspective on an Agent's Mental States and Interaction with its Environment. In: C. Castelfranchi and L. Johnson(eds.), *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'02*. ACM Press, 2002, pp. 865-872. Extended abstract in: *Proc. of the Third International Conference on Cognitive Science, ICCS 2001*, USTC Press, Beijing, 2001, pp. 551-556.

Konolige, K., and Pollack, M.E. (1989). Ascribing plans to agents: Preliminary report. In *Proc. of the 11th International Joint Conference on Artificial Intelligence*, 1989, pp. 924-930.

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

Marx, M., and Treur, J. (2001). Trust Dynamics Formalised in Temporal Logic. In: *Proc. of the Third International Conference on Cognitive Science, ICCS 2001*. USTC Press, Beijing, 2001, pp. 359-363.

Pollack, M.E. (1992). The uses of plans. *Artificial Intelligence*, 57(1), 1992, pp. 43-68.

Purang, K., Purushothaman, D., Traum, D., Andersen, C., and Perlis, D. (1999). Practical Reasoning and Plan Execution with Active Logic. In: *Proc. of the IJCAI'99 Workshop on Practical Reasoning and Rationality*.

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

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

Rosenschein, S., and Kaelbling, L.P. (1986). The Synthesis of Digital Machines with Provable Epistemic Properties. In: J.Y. Halpern (ed.), *Proc. of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge (TARK'86)*, Morgan Kaufmann, 1986, pp. 83-98.

Sosa, E., and Tooley, M. (eds.) (1993). *Causation*. Oxford University Press, 1993.

Wooldridge, M.J., and Lomuscio, A. (2001). Reasoning about Visibility, Perception and Knowledge. In: Jennings, N.R., and Lespérance, Y. (eds.), *Intelligent Agents VI, Proc. ATAL'99*. Lecture Notes in AI, Springer Verlag, 2000.