

Temporal Requirements for Anticipatory Reasoning about Intentional Dynamics in Social Contexts

Catholijn M. Jonker, Jan Treur, Wieke de Vries¹

Department of Artificial Intelligence, Vrije Universiteit Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
Email: {jonker, treur}@cs.vu.nl, URL: <http://www.cs.vu.nl/~jonker, ~treur>

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

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 et al., 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 *design stance* and the *intentional stance* are put forward. Different description levels with ontologies for emerging patterns in the simulation environment Life are used to explain the advantage of explanations using a higher level ontology; cf. [Dennett, 1987], pp. 37-39; [Dennett, 1991], pp. 37-42. In addition, Dennett uses the description levels in computer systems (actually of a chess computer), embedded (and hence visualised) in the two-dimensional Life environment as a metaphor to explain the advantage of design stance and intentional stance explanations for mental phenomena over physical stance explanations:

‘The scale of compression when one adopts the intentional stance toward the two-dimensional chess-playing computer galaxy is stupendous: it is the difference between figuring out in your head what white’s most likely (best) move is versus calculating the state of a few trillion pixels through a few hundred thousand generations. But the scale of savings is really no greater in the Life world than in our own. 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 nerve, and so forth.’ [Dennett, 1991], p. 42

In organisations, behaviour is assumed to be constrained by the organisational structure (cf. [Ferber and Gutknecht, 1998, 1999]), including, in particular, behavioural role specifications (cf. [Ferber et al., 2000]). 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

¹ Currently at: Utrecht University, Institute of Information and Computing Sciences, wieke@cs.uu.nl

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 5 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 'real world' patterns, and between the intentional notions themselves, however, are only covered partially in the literature on BDI-logics as mentioned: within a BDI-logic, for a given world state all beliefs, desires and intentions are derived at once, without internal dynamics. 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]. 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. Some environments will leave the system in that same final state, when interactions with this system are complete, and some environments will leave the system in different 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. A possible final state, then, implicitly defines, in an interactive sense, its class of environments. Dually, 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. The overall system, with its possible final states, therefore, functions as a *differentiator* of environments, with the final states implicitly defining the differentiation categories. (..) Representational content is constituted as indications of potential further interactions. (..) The claim is that such differentiated functional indications in the context of a goal-directed system constitute representation - emergent representation.'

This suggests that mental states need to be grounded in interaction histories on the one hand, and have to be related to future interactions on the other hand. However, in this literature no formalisation is proposed based on this perspective. In the formalisation introduced below, the temporal aspect of the dynamics of the interaction with the environment is made explicit and related to mental notions.

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. A temporal trace language is used to express properties on behaviour. In this temporal language, a (temporal) formula on the past in terms of the agent's input and output states defines a class of (possible) interaction histories. Formal criteria are identified that express when a (temporal) formula on the past defines a class of interaction histories that can be related to a specific belief, desire or intention. A temporal formula satisfying these criteria for a specific intentional notion is called an *external representation* or *temporal grounding* of this notion. These criteria can be used to identify a past formula that can serve as an external representation; this can be a time-consuming computational process involving the inspection of a large number of observed behaviour traces (comparable to a specific type of machine learning). However, once such an external representation formula 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) formula. The approach has been tested on operability in the implementation of 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 formal temporal trace language used in this paper is introduced. 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 formula to be a justified grounding of a specific intentional notion. In Section 4 an example application to organisation modelling is addressed. Section 5 briefly describes an implemented agent architecture for agents able to reason about the intentions of other agents. Section 6 is a discussion.

2 Formal Preliminaries

Formal ontologies (i.e., vocabularies) for the agent's input, output and internal state, are used, and formulae based on these ontologies. For simplicity, we use predicate logic to specify both ontologies and formulae. An ontology is specified as a finite set of sorts, constants within these sorts, and relations and functions over these sorts (sometimes also called a signature). The union of two ontologies is again an ontology. If functions are used, recursion is excluded, to keep the number of ground atoms finite. By considering the finite set of ground atoms as proposition symbols, the state languages can be treated as propositional.

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 constructs for observation results and communication received. The following *input atoms* are used. The observation result that ϕ holds is denoted by *observation_result*(ϕ), where ϕ is describing information on the external environment. Similarly, *communicated_by*(ϕ , C) denotes that agent C has communicated ϕ . The *agent output ontology* *OutOnt* contains constructs to represent decisions to do actions within the external world, as well as constructs for outgoing communication and observations that the agent needs to obtain. The following *output atoms* are used: *to_be_performed*(A) denotes that the agent decides to perform action A,

to_be_communicated_to(ϕ , C) means that the agent sends information ϕ to agent C, and to_be_observed(ϕ) denotes that the agent decides to perform an observation to investigate the truth of ϕ . All expressions introduced to formalise the interaction of the agent with its environment are meta-expressions; some of their arguments refer to statements in an object-level language. In the above expressions, the symbol ϕ refers to a formula based on EWOnt. The *internal agent ontology* IntOnt is used for the internal (e.g., BDI) notions. The *agent interface ontology* is defined by InterfaceOnt = InOnt \cup OutOnt; the *agent ontology* by AgOnt = InOnt \cup IntOnt \cup OutOnt, and the overall ontology by OvOnt = AgOnt \cup EWOnt.

The overall signature based on the ontologies defined above is called the *state language* (abbreviated as SL) and its formulae are called *state formulae*. By SL(Ag) the restriction of SL to the agent ontology AgOnt is denoted. Although SL is propositional, quantification in formulae is allowed and is interpreted as disjunction (in case of \exists) and conjunction (in case of \forall) of all instances. All state formulae based on a certain ontology Ont constitute the set SFOR(Ont). An *information state* of the agent is an assignment of truth values {true, false, unknown} to the set of ground atoms in SL(Ag). The set of all possible information states of the agent is denoted by IS(Ag). These notions can also be restricted to some of the ontologies related to the agent.

2.2 Temporal Language

To describe behaviour of the agent, we refer to time in a formal manner. 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 \mathbf{T} is a sequence of information states $(M^t)_{t \in \mathbf{T}}$ in IS(OvOnt). 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. Given a trace \mathcal{M} of agent Ag, the information state of the input interface at time point t is denoted by $\text{state}(\mathcal{M}, t, \text{input}(\text{Ag}))$. Analogously, $\text{state}(\mathcal{M}, t, \text{output}(\text{Ag}))$ denotes the information state of the output interface of the agent at time point t , and $\text{state}(\mathcal{M}, t, \text{internal}(\text{Ag}))$ the internal information state. We can also refer to the overall information state of a system (agents and environment) at a certain moment; this is denoted by $\text{state}(\mathcal{M}, t)$. These formalised information states can be related to state formulae via the formally defined satisfaction relation \models . If $\phi \in \text{SFOR}(\text{InOnt})$, then

$$\text{state}(\mathcal{M}, t, \text{input}(\text{Ag})) \models \phi$$

denotes that ϕ is true in this state at time point t , based on the strong Kleene semantics (e.g., [Blamey, 1986]). Comparable to the approach in situation calculus, the sorted predicate logic temporal language TL is built on atoms like the one above, using the usual logical connectives and quantification (for example, over traces, time and state formulae). 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 language.

To focus on different aspects of the agent and time, we need ways to restrict traces. Restrictions have two parameters, one for the ontologies and one for the time interval. The ontology parameter indicates which parts of the agent 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. To illustrate this, the notation $\mathcal{M}_{[0, t]}^{\text{InterfaceOnt}}$ denotes the restriction of \mathcal{M} to the past up to t and to external atoms. 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:

$$\mathcal{M}_{\text{Interval}}^{\text{Ont}}(t)(a) = \begin{cases} \mathcal{M}(t)(a) & \text{if } t \in \text{Interval and } a \text{ is a ground atom of Ont} \\ \text{unknown} & \text{otherwise} \end{cases}$$

A *past formula* for \mathcal{O} and t is a temporal formula $\psi(\mathcal{O}, t)$ 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 t' a restriction of the form $t' \leq t$, or $t' < t$ is required within the formula. The set of past formulae over ontology Ont w.r.t. t is denoted by $\text{PFOR}(\text{Ont}, t)$. Note that for any past formula $\psi(\mathcal{O}, t)$ it holds:

$$\forall \mathcal{O} \in \mathcal{W} \forall t \psi(\mathcal{O}_{[0, t]}, t) \Leftrightarrow \psi(\mathcal{O}, t).$$

To express that some formula has just become true, we introduce the following notation pronounced as *just* :

$$\begin{aligned} \oplus \text{state}(\mathcal{O}, t1, \text{interface}) \models \varphi &\equiv \\ \text{state}(\mathcal{O}, t1, \text{interface}) \models \varphi \wedge \exists t2 < t1 \forall t [t2 \leq t < t1 \Rightarrow \text{state}(\mathcal{O}, t, \text{interface}) \not\models \varphi] \\ \oplus \text{state}(\mathcal{O}, t1, \text{interface}) \not\models \varphi &\equiv \\ \text{state}(\mathcal{O}, t1, \text{interface}) \not\models \varphi \wedge \exists t2 < t1 \forall t [t2 \leq t < t1 \Rightarrow \text{state}(\mathcal{O}, t, \text{interface}) \models \varphi] \end{aligned}$$

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 notions can be attributed. Formulae expressed in the temporal language defined above will be analysed on whether or not they are adequate candidates for representations of intentional notions. In particular, conditions are given that formalise when a formula represents a belief, desire or intention.

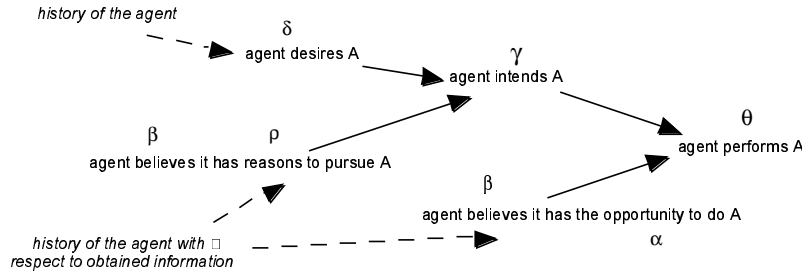


Figure 1 Relationships between the BDI notions

A basic assumption made is that an agent's internal 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, the 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 formulae that describe the history of the agent as candidates for representations of externally attributed intentional notions and actions; otherwise these formulae will not be found.

3.1 Beliefs

In the simplest approach, beliefs (β) 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: if the agent has

been informed in the past about a world fact, and no opposite information 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 was informed about this world fact (by sensing or communication), and no opposite information was received since then.

Before giving a temporal characterisation of the notion of belief an auxiliary definition is presented; let $\varphi \in \text{SFOR}(\text{EWOnt})$, then:

$$\text{Informed}(\varphi, t, \mathcal{A}) \equiv \bigoplus \text{state}(\mathcal{A}, t, \text{input}(\text{Ag})) \models \text{observation_result}(\varphi) \quad \vee \\ \exists B \in \text{AGENT} \bigoplus \text{state}(\mathcal{A}, t, \text{input}(\text{Ag})) \models \text{communicated_by}(\varphi, B)$$

Here AGENT is a sort for the agents names. So, $\text{Informed}(\varphi, t, \mathcal{A})$ means that the agent has just received information that φ is true at time point t . The following characterisation of belief is based on the assumption that an agent believes a fact if it was informed about it in the past and the fact is not contradicted by later information of the opposite. Here for φ , the *complementary formula* $\sim \varphi$ is defined as $\sim \varphi = \alpha$ if $\varphi = \neg \alpha$ and $\sim \varphi = \neg \varphi$ otherwise.

Definition (Belief Representation)

Let $\varphi \in \text{SFOR}(\text{EWOnt})$. The temporal formula $\beta(\mathcal{A}, t) \in \text{TL}$ is an externally grounded *belief formula* for φ if

$$\forall \mathcal{A} \in \mathcal{W} \forall t_1 [\beta(\mathcal{A}, t_1) \Leftrightarrow \\ \exists t_0 \leq t_1 [\text{Informed}(\varphi, t_0, \mathcal{A}) \wedge \forall t \in [t_0, t_1] \neg \text{Informed}(\sim \varphi, t, \mathcal{A})]]$$

The temporal past formula $\beta_P(\mathcal{A}, t) \in \text{PFOR}(\text{InOnt}, t)$ is called a *historical external belief representation* for φ if $\beta_P(\mathcal{A}, t)$ is an externally grounded belief formula for φ .

Note that the temporal past formula $\text{Belief}(\varphi, t_1, \mathcal{A})$ in $\text{PFOR}(\text{InOnt}, t_1)$ defined by

$$\exists t_0 \leq t_1 [\text{Informed}(\varphi, t_0, \mathcal{A}) \wedge \forall t \in [t_0, t_1] \neg \text{Informed}(\sim \varphi, t, \mathcal{A})]$$

itself is an externally grounded belief formula for φ . 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.

For the next proposition, an input atom is called *correct* with respect to the world state if

$$\forall \mathcal{A} \in \mathcal{W} \forall t [\text{Informed}(\varphi, t, \mathcal{A}) \Rightarrow \text{state}(\mathcal{A}, t, \text{EW}) \models \varphi]$$

Proposition 3.1

Let $\varphi \in \text{SFOR}(\text{EWOnt})$ be given.

a) All belief formulae for φ are temporally equivalent; i.e., if $\beta_1(\mathcal{A}, t), \beta_2(\mathcal{A}, t) \in \text{TL}$ are two externally grounded belief formulae for φ , then

$$\forall \mathcal{A} \in \mathcal{W} \forall t \quad \beta_1(\mathcal{A}, t) \Leftrightarrow \beta_2(\mathcal{A}, t)$$

b) Suppose the input atoms are correct with respect to the world state. Then at each time point there are no belief formulae true for complementary world state formulae; i.e., for any world state formula φ , if $\beta_1(\mathcal{A}, t_1)$ is a belief formula for φ and $\beta_2(\mathcal{A}, t_1)$ is a belief formula for the complementary formula $\sim \varphi$, then

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

Note that temporal equivalence of belief formulae for two internal belief representations does not imply equivalence of the two within the state logic. Only the truth values of these two formulae will be related over time, according to the dynamics of the system.

3.2 Desires and Intentions

Motivational attitudes refer in their semantics to the future actions of the agent, so it can be expected that in a characterisation a reference to future actions of the agent is made. Our assumptions are as follows. In the first place, an agent who intends to perform an action will execute the action when an opportunity (α) occurs. Moreover, the second assumption is that when an action or communication (A) is performed (θ), the agent is assumed to have *intended* (γ) to do that.

In the definition below an *action atom* $\theta(\mathcal{X}, t)$ is an atom of the form $\text{state}(\mathcal{X}, t, \text{output}(\text{Ag})) \models \psi$ with ψ an output atom: an atom of the form $\text{to_be_performed}(A)$, $\text{to_be_communicated_to}(\phi, B)$, or $\text{to_be_observed}(\phi)$.

Definition (Intention Representation)

Let $\alpha \in \text{SFOR}(\text{EWOnt})$ be an external state formula and $\theta(\mathcal{X}, t)$ an action atom. The temporal formula $\gamma(\mathcal{X}, t) \in \text{TL}$ is called an *externally grounded intention formula* for action atom $\theta(\mathcal{X}, t)$ and *opportunity* α if the following conditions are fulfilled:

Sufficiency condition for intention:

$$\begin{aligned} \forall \mathcal{X} \in \mathcal{W} \forall t_1 [\gamma(\mathcal{X}, t_1) \wedge \\ \exists t_0 \leq t_1 [\text{Informed}(\alpha, t_0, \mathcal{X}) \wedge \forall t \in [t_0, t_1] \neg \text{Informed}(\neg\alpha, t, \mathcal{X})] \\ \Rightarrow \exists t_2 \geq t_1 \theta(\mathcal{X}, t_2)] \end{aligned}$$

Necessity condition for intention:

$$\begin{aligned} \forall \mathcal{X} \in \mathcal{W} \forall t_2 [\theta(\mathcal{X}, t_2) \Rightarrow \\ \exists t_1 \leq t_2 \gamma(\mathcal{X}, t_1) \wedge \exists t_0 \leq t_1 [\text{Informed}(\alpha, t_0, \mathcal{X}) \wedge \forall t \in [t_0, t_1] \neg \text{Informed}(\neg\alpha, t, \mathcal{X})]] \end{aligned}$$

The past formula $\gamma_P(\mathcal{X}, t) \in \text{PFOR}(\text{InOnt}, t)$ is called a *historical external intention representation* for action atom $\theta(\mathcal{X}, t)$ and *opportunity* α if it is an externally grounded intention formula for $\theta(\mathcal{X}, t)$ and opportunity α .

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 formulae used for the opportunity, any belief formula (e.g., an internal belief representation) is given, then the characterisation of an intention can be reformulated.

Proposition 3.2

Let $\alpha \in \text{SFOR}(\text{EWOnt})$ an external state formula, $\beta_\alpha(\mathcal{X}, t)$ be a belief formula for α and $\theta(\mathcal{X}, t)$ an action atom. The temporal formula $\gamma(\mathcal{X}, t) \in \text{TL}$ is an externally grounded intention formula for action atom $\theta(\mathcal{X}, t)$ and opportunity α if and only if the following conditions are fulfilled:

Sufficiency condition for intention:

$$\forall \mathcal{X} \in \mathcal{W} \forall t_1 [\gamma(\mathcal{X}, t_1) \wedge \beta_\alpha(\mathcal{X}, t_1) \Rightarrow \exists t_2 \geq t_1 \theta(\mathcal{X}, t_2)]$$

Necessity condition for intention:

$$\forall \mathcal{X} \in \mathcal{W} \forall t_2 [\theta(\mathcal{X}, t_2) \Rightarrow \exists t_1 \leq t_2 \gamma(\mathcal{X}, t_1) \wedge \beta_\alpha(\mathcal{X}, t_1)]$$

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 trace	time point 0	time point 1	time point 2	time point 3	time point 4	time point 5
trace 1	food screen	no food no screen	food screen	food no screen	food no screen eat	food no screen eat
trace 2	no food no screen	food no screen	no food screen	food no screen	food no screen	food no screen eat
trace 3	no food no screen	no food no screen	food screen	food no screen	food no screen	food no screen
trace 4	food no screen	no food no screen	food screen	food screen	food no screen	food no screen eat

Table 1 Example set of observed traces

For the state formula \neg screen as opportunity, the following past formula was found to be an adequate intention representation:

$$\begin{aligned} \gamma_P(\mathcal{M}, t) = & \text{state}(s, \mathcal{M}, t, \text{input}(\text{agent})) \models \text{observed}(\text{food}) \wedge \\ & \exists t_1 \leq t \ [\text{state}(s, \mathcal{M}, t_1, \text{input}(\text{agent})) \models \text{observed}(\neg \text{food}) \wedge \\ & \exists t_2 \leq t_1 \ \text{state}(s, \mathcal{M}, t_2, \text{input}(\text{agent})) \models \text{observed}(\text{food})] \end{aligned}$$

Informally this formula 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.

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 (δ), for each relevant action there is an additional reason (ρ), so that if both the desire is present and the agent believes the reason, then the intention to perform the action will be generated. The second assumption formalised in the definition below is that every intention is based on a desire (δ), i.e., no intentions without desires. Desire representations are defined as follows:

Definition (Desire Representation)

Let an external state formula $\rho \in \text{SFOR}(\text{EWOnt})$ and an intention formula $\gamma(\mathcal{M}, t)$ be given. The temporal formula $\delta(\mathcal{M}, t) \in \text{TL}$ is called an *externally grounded desire formula* for intention $\gamma(\mathcal{M}, t)$ and *reason* ρ if the following conditions are fulfilled:

Sufficiency condition for desire:

$$\begin{aligned} \forall \mathcal{M} \in \mathcal{W} \ \forall t_1 \ [\delta(\mathcal{M}, t_1) \wedge \\ \exists t_0 \leq t_1 \ [\text{Informed}(\rho, t_0, \mathcal{M}) \wedge \forall t \in [t_0, t_1] \ \neg \text{Informed}(\neg \rho, t, \mathcal{M})] \\ \Rightarrow \exists t_2 \geq t_1 \ \gamma(\mathcal{M}, t_2)] \end{aligned}$$

Necessity condition for desire:

$$\begin{aligned} \forall \mathcal{M} \in \mathcal{W} \ \forall t_2 \ [\gamma(\mathcal{M}, t_2) \Rightarrow \\ \exists t_1 \leq t_2 \ \delta(\mathcal{M}, t_1) \wedge \exists t_0 \leq t_1 \ [\text{Informed}(\rho, t_0, \mathcal{M}) \wedge \forall t \in [t_0, t_1] \ \neg \text{Informed}(\neg \rho, t, \mathcal{M})]] \end{aligned}$$

The past formula $\delta_P(\mathcal{M}, t) \in \text{PFOR}(\text{InOnt}, t)$ is called a *historical external desire representation* for intention $\gamma(\mathcal{M}, t)$ and (additional) *reason* ρ if it is an externally grounded desire formula for intention $\gamma(\mathcal{M}, t)$ and reason ρ .

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 formulae used for the reason, any belief formula (e.g., an internal belief representation) is given, then the characterisation of a desire can be reformulated.

Proposition 3.3

Let $\rho \in \text{SFOR}(\text{EWOnt})$ be an external state formula, $\beta_\rho(\mathcal{M}, t)$ a belief formula for ρ and $\gamma(\mathcal{M}, t)$ an intention formula. The temporal formula $\delta(\mathcal{M}, t) \in \text{TL}$ is an externally grounded desire formula for intention $\gamma(\mathcal{M}, t)$ and reason ρ if and only if the following conditions are fulfilled:

Sufficiency condition for desire:

$$\forall \mathcal{M} \in \mathcal{M} \forall t_1 [\delta(\mathcal{M}, t_1) \wedge \beta_\rho(\mathcal{M}, t_1) \Rightarrow \exists t_2 \geq t_1 \gamma(\mathcal{M}, t_2)]$$

Necessity condition for desire:

$$\forall \mathcal{M} \in \mathcal{M} \forall t_2 [\gamma(\mathcal{M}, t_2) \Rightarrow \exists t_1 \leq t_2 \delta(\mathcal{M}, t_1) \wedge \beta_\rho(\mathcal{M}, t_1)]$$

4 Anticipatory Reasoning and Acting in Organisations

Viewed from a dynamic perspective, organisational structure (cf. [Ferber and Gutknecht, 1998]), provides specifications of constraints on role behaviour and interactions (cf. [Ferber and Gutknecht, 1999; Ferber et al., 2000]). 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 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 reverse of the medal, however, is that this freedom also may provide possibilities to agents to avoid (based on their individual interest) certain behaviours as expected by others, and thus may 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:
 $\text{observation_result}(\text{task_urgent})$, $\text{observation_result}(\text{own_agenda_full})$,
 $\text{observation_result}(\text{colleagues_agenda_not_full})$,
 $\text{observation_result}(\text{colleagues_capable_of_task})$,
 $\text{communicated}(\text{task_request})$, $\text{communicated}(\text{task_possibility})$
- The *output ontology* OutOnt includes $\text{tbc}(\text{task_rejection})$. Here tbc is short for 'to be communicated'.

Define the past formula $\delta_p(\mathcal{M}, t) \in \text{PFOR}(\text{Ont}, t)$ for the *desire* to avoid the task by

$$\begin{aligned} \text{state}(\mathcal{M}, t, \text{input}(\text{Ag})) & \models \text{communicated}(\text{task_request}) \wedge \\ \text{state}(\mathcal{M}, t, \text{input}(\text{Ag})) & \models \text{observation_result}(\text{task_urgent}) \wedge \\ \neg \exists t_0 < t & \text{state}(\mathcal{M}, t_0, \text{input}(\text{Ag})) \models \text{communicated}(\text{task_possibility}) \end{aligned}$$

The *reason* ρ to generate an avoidance intention is:

$$\text{colleagues_agenda_not_full} \wedge \text{colleagues_capable_of_task}$$

The past formula $\gamma_p(\mathcal{M}, t) \in \text{PFOR}(\text{Ont}, t)$ for the *intention* to avoid the task is defined in short form by

$$\delta_p(\mathcal{M}, t) \ \& \ \text{Belief}(\text{colleagues_agenda_not_full} \wedge \text{colleagues_capable_of_task}, t, \mathcal{M})$$

The extensive form is

$$\begin{aligned} \text{state}(\mathcal{M}, t, \text{input}(\text{Ag})) & \models \text{communicated}(\text{task_request}) \wedge \\ \text{state}(\mathcal{M}, t, \text{input}(\text{Ag})) & \models \text{observation_result}(\text{task_urgent}) \wedge \\ \neg \exists t_0 < t & \text{state}(\mathcal{M}, t_0, \text{input}(\text{Ag})) \models \text{communicated}(\text{task_possibility}) \wedge \\ \exists t_0 \leq t_1 & [\text{Informed}(\text{colleagues_agenda_not_full} \wedge \text{colleagues_capable_of_task}, t_0, \mathcal{M}) \wedge \\ \forall t \in [t_0, t_1] & \neg \text{Informed}(\sim(\text{colleagues_agenda_not_full} \wedge \text{colleagues_capable_of_task}), t, \mathcal{M}) \end{aligned}$$

The *opportunity* α to perform the avoidance action is:

$$\text{own_agenda_full}$$

The past formula $\theta_p(\mathcal{A}, t) \in \text{PFOR}(\text{Ont}, t)$ for the *action* to avoid the task is defined in its short form by

$$\gamma_p(\mathcal{A}, t) \ \& \ \text{Belief}(\text{own_agenda_full}, t, \mathcal{A})$$

Given this formalisation 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 formula 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 An Agent Architecture for Intention Attribution

In Section 3 a formal analysis was made of possible temporal representations and criteria to be used for (externally attributed) intentional notions to model other agents' behaviour. In this section based on these criteria and representations, an agent architecture is introduced for agents that are capable of anticipatory reasoning about the occurrence of intentional notions on the basis of observed behaviour of other agents. In particular, this agent architecture is capable of:

- observation of other agents' behaviour
- maintenance of a history of behavioural traces of other agents
- derivation of adequate temporal intentional representations to model other agents' behaviour on the basis of observed behaviour
- explanation of observed behaviour using these temporal intentional representations
- analysis of the circumstances related to the occurrence of intentions preceding actual behaviour of other agents on the basis of the intentional representations

The compositional Generic Agent Model GAM described in [Brazier et al., 2000] was used as a starting point for the design of this agent architecture. In particular, the following components are relevant, and were specialised and instantiated:

- Own Process Control, where the decision is made on which other agents to focus the monitoring,
- World Interaction Management, where the observation of other agents is managed,
- Maintenance of Agent Information, where agent models are created and maintained.

5.1 Observation of other agents' behaviour

At the top level a multi-agent system consists of a number of agents, communicating to each other, and an External World component. Both the execution of actions in the world and observation by an agent are realized by an interaction of the agent with the External World component. Within the External World component, among others, the behaviour of all agents in terms of the observations and actions they perform is represented. Two types of observation are possible: (1) *passive*: the agent receives observation information from the External World without making any decision to do so, and (2) *active*: the agent decides to observe (to focus on) certain aspects of the world and only receives observation information of these aspects. In particular, if

other agents' behaviours are observed, this can be done passively, unfocussed (all behaviour of other agents is monitored), or actively, focussed (a selected set of agents is monitored).

In the first, passive case, observation information on other agents' behaviour is received (from the External World) by the agent, within the agent transferred to the component World Interaction Management, where it is identified as agent information; subsequently it is transferred to the component Maintenance of Agent Information.

In the second, active case, bidirectional interaction of the agent with the External World is performed. First the agent decides which other agents to monitor (within its component Own Process Control), and this focus information is transferred to World Interaction Management, where (persistent) observation foci are generated. These observation foci are transferred to the output of the agent, and from there to the External World component. The External World transfers the required observation results to the agent, where the received observation information is treated in the same manner as in the passive case.

In both cases the External World provides observation results to the monitoring agent about which observations have been performed by the monitored agents, and which actions they performed. The model can easily be extended to include the monitoring of communication as well.

5.2 Maintenance of Agent Information

The component Maintenance of Agent Information is composed of four sub-components (see Figure 2):

- Maintenance of Agent Histories
- Intentional Representation Determination
- Maintenance of Agent States
- Behaviour Analysis

Within the component Maintenance of Agent Histories, the received observations are labeled by time points and stored as traces.

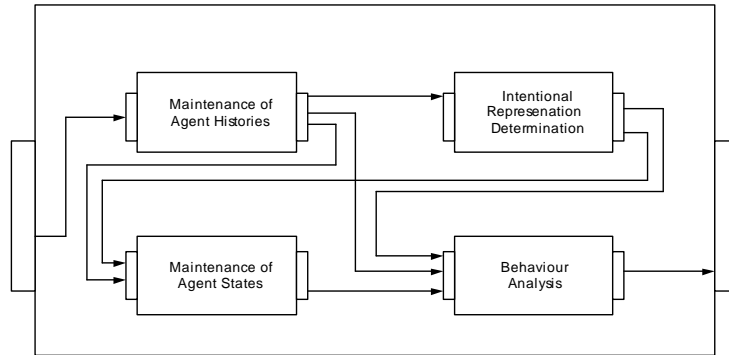


Figure 2 Composition of the component Maintenance of Agent Information

Within the component Intentional Representation Determination temporal formulae are identified that satisfy the criteria for intentional notions formulated in Section 3 (this is a computation-intensive process that can be performed off-line, i.e. during idle times). At each time point the component Maintenance of Agent States maintains a model of the other agents state in terms of the (attributed) intentions they have. This

information is used in the component Behaviour Analysis to perform anticipatory deliberation on which (histories of) circumstances may lead to the occurrence of which intentions, or to analyse observed behaviour in retrospect.

The content of component Intentional Representation Determination consists of an implementation (in Prolog) of a formula generator to generate candidate temporal past formulae for intentional notions, and a formula checker which verifies whether the criteria expressed in Section 3 are satisfied by a given candidate formula for the given behaviour traces from history. If indeed a candidate formula fulfils these criteria, and has the lowest complexity (in nesting of logical connectives), then this formula is selected as an adequate representation of the intentional notion, and transferred to the component Maintenance of Agent States.

5.3 Prolog Representations

Temporal formulae are represented by nested term structures using the logical connectives; e.g., the example formula $\gamma(\mathcal{M}, t)$ from Section 3 is represented by

```
and(holds(state(M, T, input(agent)), observed(food), true),
    ex(T1≤T, and(holds(state(M, T1, input(agent)), observed(neg(food)), true),
        ex(T2≤T1, holds(state(M, T2, input(agent)), observed(food), true))) ) ).
```

Traces are represented by Prolog facts of the form, e.g.,

```
holds(state(m1, t(2), input(agent)), observed(neg(food))), true).
```

Where $m1$ is the trace name, $t(2)$ time point 2; it is indicated that state formula $\text{informed}(\text{neg}(a))$ is true in the system's state at time point 2. One Prolog programme used within component Identification of Intentional Representations generates these term structures for temporal formulae by iterative deepening of their nesting, starting at depth 0. Another Prolog programme used within this component performs temporal formula checking using Prolog rules such as:

```
sat(and(F,G)) :- sat(F), sat(G).
```

These rules reduce complex temporal formulae to their constituting state atoms and evaluates these at specific time points of the give trace(s). In this process, to be able to work with abstraction to manage complexity, it is possible to introduce a new formula name F to denote a more complex formula G by $\text{denotes}(F, G)$; e.g.,

```
denotes(belief(M, T, C, pos),
    ex(T1≤T, and(observed(M, T1, C, pos), all(T2≤T, implies(le(T1, T2), neg(observed(M,
    T2, C, neg))))))).
```

where

```
denotes(observed(M, T, C, S),
    holds(state(M, T, input(agent)), observed(C), S)).
```

Moreover, the formula $\text{initiates}(M, T, C)$ denoting action initiation is defined by:

```
denotes(initiates(M, T, C),
    holds(state(M, T, output(agent)), to_be_performed(C), true)).
```

The sufficiency condition for a temporal formula $\gamma(M, T1)$ to represent an intention then is expressed as

```
all(T1, implies(and(gamma(M, T1), belief(M, T1, b, pos) ), ex(T2≥T1, initiates(M, T2, c))))).
```

Similarly the other notions are represented.

6 Discussion

The formal analysis and implementation presented in this paper differs from the approaches 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.

The temporal trace language used in our approach is much more expressive than standard temporal logics in a number of respects. 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, in our approach states are *three-valued*; the standard temporal logics are based on two-valued states, which implies that for a given trace a form of closed world assumption is imposed. This means that, for example, in Concurrent MetateM (cf., [Fisher, 1994]), if the executable temporal logic specification leaves some atom unspecified, during construction of a trace the semantics will force it to be false. To avoid this, an atom has to be split into a positive and a negative variant.

Fourth, the possibility to quantify over traces allows for specification of *more complex behaviours*. As within most temporal logics, reactiveness and pro-activeness properties be specified. In addition, in our language also properties expressing different types of adaptive behaviour can be expressed. For example a property such as 'exercise improves skill', which is relative in the sense that it involves the comparison of two alternatives for the history. This type of property can be expressed in our language, whereas in standard forms of temporal logic different alternative histories cannot be compared. Fifth, in our language it is possible to define *local languages for parts* of a system. For example, the distinction between internal, external and interface languages is crucial, and is supported by the language, which also entails the possibility to quantify over system parts; this allows for specification of system modification over time. Sixth, since state properties are used as first class citizens in the temporal trace language, it is possible to explicitly refer to them, and to quantify over them, enabling the specification of what are sometimes called *second-order properties*, which are used in part of the philosophical literature (e.g., [Kim, 1998]) to express functional roles related to mental properties or states.

In the current paper only part of the features of the language as discussed above are exploited. Due to the simplifying assumptions on the intentional notions addressed here, for this focus the job could also be done by a less expressive language. However, then the approach is less generic and will not be extendable to more complex behaviours and mental properties, such as, for example, relative adaptive behaviours. The language used is meant to support a more *generic* perspective and anticipates on these types of more complex behaviours which are in the focus of our further research. As an 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 our temporal trace language.

The application of our formalisation of intentional dynamics in anticipatory reasoning (and acting) makes it quite useful to analyse certain phenomena often occurring in organisations. As shown by a number of examples in Section 4, usually organisations leave some freedom in performing a certain role. To cooperate with other agents with such freedom, agents within an organisation not seldomly try to affect, in an anticipatory manner, the circumstances that may lead to the generation of other agent's beliefs, desires and intentions. The capability of performing such anticipatory reasoning and acting may be crucial within organisations (where often some of the agents have certain 'directions for use'). To avoid unnecessary obstruction of the organisation's processes, these 'directions of use' better can be taken into account in cooperation. Section 4 shows in detail how this can be done based on the framework introduced in Section 3. Based on a number of experiences (observed

traces) a temporal representation can be identified; this may be a computationally expensive process which has to be performed once (for example, off-line). After such a representation has been identified, it can be reused in a very efficient manner in all relevant new situations the agent encounters (on-line). The formal analysis can be supported by software as discussed in Section 5. This software, which was not yet optimized, suffices to automatically generate and verify temporal representations up to a depth of 4 in nesting of logical connectives, which covers already many interesting and nontrivial examples, including the examples in Sections 3 and 4 above. Further work will be undertaken to improve efficiency by replacing the unfocussed generation of arbitrary temporal formulae by a more intelligent process.

Our approach has its perspective on grounding of mental states in the interaction in common with [Bickhard, 1993; Christensen and Hooker, 2001]: in particular on the relation between internal agent state and interaction with the environment in the past, and potential further interactions in the future (see the citation in the introduction above). Also in [Clark, 1997] emphasis is put on the functioning of cognition in interaction with the environment. A difference is that in our approach a formalisation is proposed, and that an explicit relation of the interaction patterns with a number of wellknown intentional (BDI) notions is addressed. Other differences are that in our approach no commitments to specific internal (goal-directed) system structures and specific internal states need to be made.

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 ϕ if ϕ 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, 2000], which concentrates on the informational aspects, and abstracts from motivational and temporal aspects; actually, in [Wooldridge, 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.

From an application-oriented perspective, by means of the implementation of a dedicated agent architecture it has been shown that the defined notions and criteria are operational and provide a basis to develop applications of agents that monitor and interpret the behaviour of other agents. In research on plan recognition, such as [Allen, 1983; Konolige and Pollack, 1989], 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 formulas representing the beliefs, desires and intentions of this agent. Once such a past formula has been found this makes it possible again and again to anticipate on the generation of intentions at forehand, without any action being performed by the actor. These formulas are tested against information on the behaviour of the observed agent during a significant number of (possible) interactions of the agent. No domain knowledge on actions and effects is used.

References

- [Allen, 1983] J.F. Allen. Recognizing intentions from natural language utterances. In M. Brady and R.C. Berwick, eds., *Computational Models of Discourse*. MIT Press, Cambridge, Ma., 1983.
- [Bickhard, 1993] Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, pp. 285-333.
- [Blamey, 1986] S. Blamey, Partial Logic, in: D. Gabbay and F. Guenther (eds.), *Handbook of Philosophical Logic*, Vol. III, 1-70, Reidel, Dordrecht, 1986.
- [Brazier et al., 2000] Brazier, F.M.T., Jonker, C.M., and Treur, J., Compositional Design and Reuse of a Generic Agent Model. *Applied Artificial Intelligence Journal*, vol. 14, 2000, pp. 491-538.
- [Christensen and Hooker, 2001] Christensen, W.D. and C.A. Hooker (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.
- [Clark, 1997] Clark, A. *Being There: Putting Brain, Body and World Together Again*. MIT Press, 1997.
- [Cohen and Levesque, 1990] Cohen, P.R. and Levesque, H.J. (1990). Intention is Choice with Commitment. *Artificial Intelligence* vol. 42 (1990), pp. 213-261.
- [Dennett, 1987] Dennett, D.C. (1987). *The Intentional Stance*. MIT Press. Cambridge Mass, 1987.
- [Dennett, 1991] Dennett, D.C. (1991). Real Patterns. *The Journal of Philosophy*, vol. 88, 1991, pp. 27-51.
- [Ferber and Gutknecht, 1998] Ferber, J. and Gutknecht, O. , 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 and Gutknecht, 1999] Ferber, J. and Gutknecht, O. , Operational Semantics of a role-based agent architecture. In: Jennings, N.R., Lesperance, Y. (eds.). *Intelligent Agents VI. Proceedings of the 6th Int. Workshop on Agent Theories, Architectures and Languages*. Lecture Notes in AI, vol. 1757, Springer Verlag, 1999.
- [Ferber et al., 2000] Ferber, J., Gutknecht, O., Jonker, C.M., Mueller, J.P., and Treur, J., Organization Models and Behavioural Requirements Specification for Multi-Agent Systems (extended abstract). In: *Proc. of the Fourth International Conference on Multi-Agent Systems, ICMAS 2000*. IEEE Computer Society Press, 2000. Extended version in this volume.
- [Fisher, 1994] Fisher, M., 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, pp. 480–505.
- [Jonker and Treur, 1999] 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, 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.
- [Kim, 1998] Kim, J. *Mind in a Physical world: an Essay on the Mind-Body Problem and Mental Causation*. MIT Press, Cambridge, Mass.
- [Konolige and Pollack, 1989] K. Konolige and M.E. Pollack. Ascribing plans to agents: Preliminary report. In *Proc. of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989, pp. 924-930.
- [Linder et al., 1996] 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, pp. 17-32.
- [Pollack, 1992] M.E. Pollack. The uses of plans. *Artificial Intelligence*, 57(1), pp. 43-68, 1992.
- [Port and van Gelder, 1995] Port, R.F., Gelder, T. van (eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, Mass.
- [Rao and Georgeff, 1991] Rao, A.S. and Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-Architecture. In: (J. Allen, R. Fikes and E. Sandewall, ed.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (KR'91), Morgan Kaufmann, 1991, pp. 473-484.
- [Rosenschein and Kaelbling, 1986] 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, pp. 83-98.
- [Wooldridge, 2000] Wooldridge, M.J. (2000). Reasoning about Visibility, Perception and Knowledge. In: Jennings, N.R., and Lesperance, Y. (eds.), *Intelligent Agents VI, Proc. ATAL'99*. Lecture Notes in AI, Springer Verlag, 2000.