A Two-Level BDI-Agent Model for Theory of Mind and its Use in Social Manipulation

Tibor Bosse, Zulfiqar A. Memon, Jan Treur[1]

Abstract. This paper introduces a formal BDI-based agent model for Theory of Mind. The model uses BDI-concepts to describe the reasoning process of an agent that reasons about the reasoning process of another agent, which is also based on BDI-concepts. A case study illustrates how the model can be used for social manipulation. This case study addresses the scenario of a manager that reasons about the task avoiding behaviour of his employee. For this scenario, a number of simulation experiments have been performed, and some of their results are discussed.

1 INTRODUCTION

To function efficiently in social life and within organisations, it is useful if agents can reason about the actual and potential behaviour of the agents around them. To this end, it is very helpful for these agents to have capabilities to predict in which circumstances other agents will show certain appropriate or inappropriate behaviours. If for a considered other agent, generation of actions is assumed to be based on a BDI-model, prediction of such actions will involve reasoning based on a Theory of Mind [3], [5], [16] involving beliefs, desires and intentions as a basis for the behaviour. Reasoning based on such a Theory of Mind can be exploited in two different manners. The first manner is just to predict the behaviour in advance, in order to be prepared that it will occur (*social anticipation*). For example, if an agent B has done things that are known as absolutely unacceptable for an organisation (or a relationship), then he or she may be able to predict and therefore be prepared on what will happen after a manager (or partner) agent A learns about it.

A second manner to exploit reasoning based on a Theory of Mind is to try to affect the occurrence of certain beliefs, desires and intentions at forehand, by manipulating the occurrence of circumstances that are likely to lead to them (*social manipulation*). For example, the agent B just mentioned can try to hide facts so that the manager (or partner) agent A will never learn about the issue. Such capabilities of anticipatory and manipulatory reasoning based on a Theory of Mind about the behaviour of colleague agents are considered quite important, not to say essential, to function smoothly in social life.

This type of reasoning has an information acquisition and analysis aspect, and a preparation and action aspect. To describe the latter aspect, for the agent using a Theory of Mind, a model for action preparation based on beliefs, desires and intentions can be used as well. For example, for agent B discussed above, the desire can be generated that agent A will not perform the action to fire (or break up with) him or her, and that agent A will in particular not generate the desire or intention to do so. Based on this desire, the

[1] Vrije Universiteit Amsterdam, Department of Artificial Intelligence. De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. URL: http://www.few.vu.nl/~{tbosse, zamemon, treur}. E-mail: {tbosse, zamemon, treur}@few.vu.nl.

refined desire can be generated that agent A will not learn about the issue. Based on the latter desire, an intention and action can be generated to hide facts for agent A. Notice that agent B reasons on the basis of BDI-models at two different levels, one for B itself, and one as the basis for the Theory of Mind to reason about agent A. It is this two-level architecture that is worked out in this paper in a computational model.

The modelling approach used for this computational model is based on the modelling language LEADSTO [7]. In this language, direct temporal dependencies between two state properties in successive states are modelled by *executable dynamic properties*. The LEADSTO format is defined as follows. Let α and β be state properties of the form 'conjunction of ground atoms or negations of ground atoms'. In the LEADSTO language the notation $\alpha \rightarrow_{e, f, g, h} \beta$, means:

> *If state property α holds for a certain time interval with duration g, then after some delay (between e and f) state property β will hold for a certain time interval of length h.*

Here, atomic state properties can have a qualitative, logical format, such as an expression desire(d), expressing that desire d occurs, or a quantitative, numerical format such as an expression has_value(x, v) which expresses that variable x has value v.

In Section 2, first the general BDI-model is explained. In Section 3, this BDI-model is illustrated by a case study about an employee that shows task avoiding behaviour. Next, Section 4 describes how the simple model can be extended to a BDI-model of an agent that reasons about another agent's BDI-model and uses this for social manipulation. In Section 5, this two-level BDI-model is illustrated by a case study that elaborates upon the example addressed in Section 3. This case study addresses the scenario of a manager that reasons about the task avoiding behaviour of his employee, and how to prevent that behaviour. Based on this model, some simulation experiments and their results are presented in Section 6. Section 7 discusses related work, and Section 8 concludes the paper with a discussion.

2 THE BDI-MODEL

The BDI-model bases the preparation and performing of actions on beliefs, desires and intentions (e.g., [11], [14], [18]). This model shows a long tradition in the literature, going back to Aristotle's analysis of how humans (and animals) can come to actions; cf. [1], [2]. He discusses how the occurrence of certain internal (mental) state properties within the living being entail or cause the occurrence of an action in the external world. Such internal state properties are sometimes called by him 'things in the soul', for example, sensation, reason and desire:

> 'Now there are three things in the soul which control action and truth - sensation, reason, desire.' [1], Book VI, Part 2.

Here, sensation indicates the sensing of the environment by the agent, which leads, (in modern terms) to internal representations, called beliefs. Reason indicates the (rational) choice of an action that is reasonable to fulfil the given desire. Based on this, Aristotle introduced the following pattern to explain action (called practical syllogism):

> If   A has a desire D
> and   A has the belief that X is a (or: the best) means to achieve D
> then   A will do X

The BDI-model incorporates such a pattern of reasoning to explain behaviour in a refined form. Instead of a process from desire to action in one step, as an intermediate stage first an intention is generated, and from the intention the action is generated. Thus the process is refined into a two-step process. See Figure 1 for the generic structure of the BDI-model in causal-graph-like style, as often used to visualise LEADSTO specifications. Here the box indicates the borders of the agent, the circles denote state properties, and the arrows indicate dynamic properties expressing that one state property leads to (or causes) another state property. In this model, an action is performed when the subject has the intention to do this action and it has the belief that certain circumstances in the world are fulfilled such that the opportunity to do the action is there. Beliefs are created on the basis of observations. The intention to do a specific type of action is created if there is some desire D, and there is the belief that certain circumstances in the world state are there, that make it possible that performing this action will fulfil this desire (this is the kind of rationality criterion discussed above; e.g., what is called means-end analysis is covered by this). Whether or not a given action is adequate to fulfil a given desire depends on the current world state; therefore this belief may depend on other beliefs about the world state. Instantiated relations within the general BDI-model as depicted by arrows in graphical format in Figure 1 can be specified in formal LEADSTO format as follows:

> desire(D) $\wedge$ belief(B1) $\rightarrow$ intention(P)
> intention(P) $\wedge$ belief(B2) $\rightarrow$ performs(P)

with appropriate desire D, action P and beliefs B1, B2. Note that the beliefs used here both depend on observations, as shown in Figure 1. Furthermore, $\wedge$ stands for the conjunction operator (and) between the atomic state properties (in the graphical format denoted by an arc connecting two (or more) arrows). Often, dynamic properties in LEADSTO are presented in *semi-formal* format, as follows:

> At any point in time
> if   desire D is present
>   and   the belief B1 is present
> then   the intention for action P will occur
>
> At any point in time
> if   the intention for action P is present
>   and   the belief B2 is present
> then   the action P will be performed

As a generic template, including a reference to the agent X concerned, this can be expressed by:

For any desire D, world state property Z, and action Y such that has_reason_for(X, D, Z, Y) holds:

> desire(X, D) $\wedge$ belief(X, Z) $\twoheadrightarrow$ intention(X, Y)

For any world state property Z and action Y such that is_opportunity_for(X, Z, Y) holds:

> intention(X, Y) $\wedge$ belief(X, Z) $\twoheadrightarrow$ performs(X, Y)

Here has_reason_for(X, D, Z, Y) is a relation that can be used to specify which state property Z is considered a reason to choose a certain intention Y for desire D. Similarly is_opportunity_for(X, Z, Y) is a relation that can be used to specify which state property Z is considered an opportunity to actually perform an intended action Y.

Assuming that beliefs are available, what remains to be generated in this model are the desires. For desires, there is no generic way (known) in which they are to be generated in the standard model. Often, in applications, generation of desires depends on domain-specific knowledge.

3 A BDI-MODEL FOR TASK AVOIDANCE

To illustrate the BDI-model described above by a specific example, the following scenario is addressed (in the domain of an organisation); notice that here no Theory of Mind is involved.

Task Avoidance Case
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, and he believes that capable colleagues are available with not 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.'*

Below, this example is formalised, using the BDI-model as introduced above. First, only the behaviour of the employee is addressed (in Section 5, the deliberation process of the manager is addressed as well). To this end, the example is made more precise as follows:

The *desire* to avoid a task is created after time t by the employee if the following holds at time t:

- the employee has the belief that a task is requested that has to be finished soon
- the employee has the belief that he did not hear of the possibility that the task may come at any earlier time point

The *intention* to avoid a task is generated after time t if the following holds at time t:

- the desire to avoid the task is available
- the belief that capable colleagues are available (not full booked)

The *action* to avoid the task is generated after time t if the following holds at time t:

- the intention to avoid the task is available
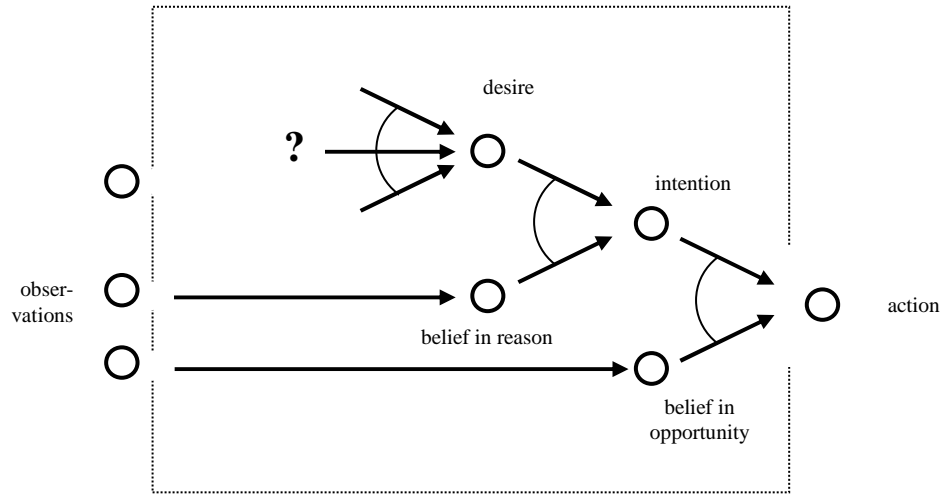- the belief that the employee's own agenda is full

Figure 1. Structure of the general BDI-model

Using the generic template discussed at the end of Section 2, via the relations

    has_reason_for(A, lower_workload, capable_colleagues_available,
                                                    avoid_task)
    is_opportunity_for(A, own_agenda_full, avoid_task)

the following model for agent A is obtained:

    belief(A, task_may_come) ∧ belief(last_minute_request) →
    desire(A, lower_workload)

    desire(A, lower_workload) ∧ belief(A, capable_colleagues_available) →
    intention(A, avoid_task)

    intention(A, avoid_task) ∧ belief(A, own_agenda_full) →
    performs(A, avoid_task)

4 THE TWO-LEVEL BDI-MODEL

As an instance of the *instrumentalist perspective* and opposed to explanations from a direct physical perspective (the physical stance), in [9], [10], the *intentional stance* (or folk-psychological stance) is put forward. In [10], Dennett explains the advantage 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 photons from brick to eyeball, the neurotransmitters from optic nerve to motor nerve, and so forth.' [10], p.42

According to the intentional stance, an agent is assumed to decide to act and communicate based on intentional notions such as 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 interaction (i.e., by observation and communication), but also on information acquired by interaction 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 a basis; cf. [10].

In the model presented in this paper, the instrumentalist perspective is taken as a point of departure for a Theory of Mind. More specifically, the model describes the reasoning process of an agent B that applies the intentional stance to another agent A by attributing beliefs, desires en intentions. Thus, for agent B a Theory of Mind is obtained using concepts for agent A's beliefs, desires and intentions. For example, in case a manager has an important last-minute task for his employee, but he knows that this employee often shows avoidance behaviour for last-minute tasks, he may analyse in more detail under which circumstances the employee may generate the desire and intention to avoid this task, and the related beliefs in reason and opportunity.

As a next step, the model is extended with BDI-concepts for agent B's own beliefs, desires and intentions as well. By doing this, agent B is able to not only *have* a theory about the mind of agent A, but also to *use* it within its own BDI-based reasoning processes. To this end, a number of meta-representations expressed by meta-predicates are introduced, e.g.:

    belief(B, desire(A, D))

This expresses that agent B believes that agent A has desire D.

    desire(B, not(intention(A, X)))

This expresses that agent B desires that agent A does not intend action X.

    belief(B, depends_on(performs(A, X), intention(A, X)))

This expresses that agent B believes that, whether A will perform action X depends on whether A intends to do X. Note that the third meta-statement has a more complex structure than the other two, since it represents a statement about a *dynamic property*, rather than a statement about a *state property*. These dependencies can be read from a graph such as depicted in Figures 1 and 2 (right hand side). For example, it is assumed that agent B knows part of this graph in his Theory of Mind, expressed by beliefs such as:

    belief(B, depends_on(performs(A, X), intention(A, X)))
    belief(B, depends_on(performs(A, P), belief(A, B2)))
    belief(B, depends_on(intention(A, P), desire(A, D)))
    belief(B, depends_on(intention(A, P), belief(A, B1)))

belief(B, depends_on(desire(A, D), belief(A, B3)))
belief(B, depends_on(belief(A, X), hears(A, X)))

Desire refinement in the BDI-model for an agent B attributing motivations to an agent A is formulated (in LEADSTO format) by:

desire(B, X) $\land$ belief(B, depends_on(X, Y)) $\longrightarrow\!\!\!\rightarrow$ desire(B, Y)

desire(B, not(X)) $\land$ belief(B, depends_on(X, Y)) $\longrightarrow\!\!\!\rightarrow$ desire(B, not(Y))

Moreover the following schemes for intention and action generation are included in the model. For any desire D, world state property Z, and action Y such that has_reason_for(B, D, Z, Y) holds:

desire(B, D) $\land$ belief(B, Z) $\longrightarrow\!\!\!\rightarrow$ intention(B, Y)

For any world state property Z and action Y such that is_opportunity_for(B, Z, Y) holds:

intention(B, Y) $\land$ belief(B, Z) $\longrightarrow\!\!\!\rightarrow$ performs(B, Y)

Moreover, some dynamic properties of the world are needed:

performs(B, tell(A, C)) $\longrightarrow\!\!\!\rightarrow$ holds_in_world(communication(B, A, C))

holds_in_world(communication(B, A, C)) $\longrightarrow\!\!\!\rightarrow$ hears(A, C)

For an overview of the complete two-level BDI-model, see Fig. 2.

5 A TWO-LEVEL BDI-MODEL FOR REASONING ABOUT TASK AVOIDANCE

The above model can be used to describe how the manager agent (from the case described in Section 3) can reason and act in an anticipatory manner to avoid the employee's avoidance desire, intention and/or action to occur. The initial desire of B is that A does not perform the action to avoid the task:

desire(B, not(performs(A, avoid_task)))

Fulfilment of this desire can be obtained in the following three manners:

Avoiding A's desire to occur
This can be obtained when the employee hears in advance that possibly a last minute task may occur. This will make the second condition in A's desire generation as described in Section 3 fail.

Avoiding A's intention to occur (given that the desire occurs)
This can be obtained by refutation of the belief that plays the role of the reason to generate the intention in A's intention generation as described in Section 3, e.g., when the employee hears that colleagues do not have the required expertise.

Avoiding A's action to occur (given that the intention occurs)
This can be obtained by refutation of the belief that plays the role of opportunity in A's desire action as described in Section 3, e.g., when his agenda is not full-booked.

For convenience, the model does not make a selection but addresses all three options to prevent the avoidance action. This means that B generates desires for:

- A hears about the possibility of a last-minute task in advance

    hears(A, task_may_come)

- A hears that no colleagues that are capable of performing the task are available

    hears(A, not(capable_colleagues_available))

- A hears that his agenda is not full-booked

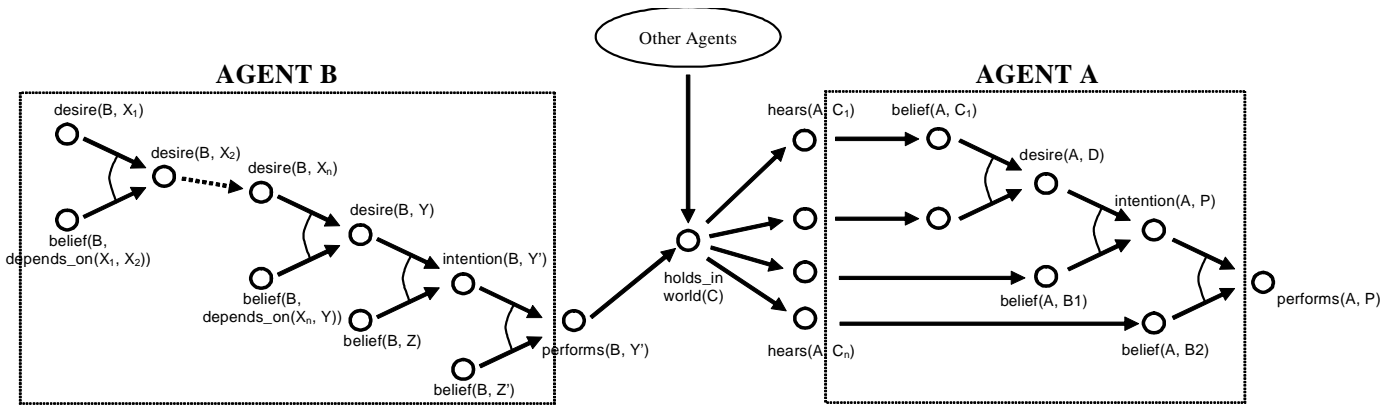   hears(A, not(own_agenda_full))

Figure 2. Structure of the Two-Level BDI-model

To fulfil these desires, intentions are to be generated by B to perform actions such as:

- B tells A about the possibility of a last-minute task in advance

    performs(B, tell(A, task_will_come))

- B tells A that no colleagues that are capable of performing the task are available

    performs(B, tell(A, not(capable_colleagues_available))

- B tells A that some of the (perhaps less interesting) tasks were taken from A's agenda and were re-allocated to a colleague

    performs(B, tell(A, not(own_agenda_full)))

Reason for B to choose for these actions is
- the belief of B that telling something will lead to the person hearing it
    belief(B, adequate_communication(B, A))

Moreover, these intentions of B can lead to the corresponding actions when the following belief of B in opportunity is there:

- the belief that A is available for B to talk to
    belief(B, available_for(A, B))

In addition to the generic BDI-model shown in Section 4, the following specific relations are used to model the case study:

    has_reason_for(B, hears(A, C), adequate_communication, tell(A, C))
    is_opportunity_for(B, available_for(A, B), tell(A, C))

Note that the last minute request itself is an event that not necessarily comes from agent B; it can come from any agent, for example a Director agent. It is modelled as an event in LEADSTO.

6 SIMULATION EXPERIMENTS

In a number of simulation experiments, the two-level BDI-model has been applied to the case study as described in Section 5. To this end, the LEADSTO software environment [7] has been used. In Figure 3 and 4, examples of resulting simulation traces are shown. In these figures, time is on the horizontal axis; the state properties are on the vertical axis. A box on top of a line indicates that a state property is true. Note that, due to space limitations, only a selection of the relevant atoms is shown. Figure 3 is the resulting simulation trace of the situation explained in Section 3 in which *no* Theory of Mind is involved, i.e., only the behavior of employee is addressed, without social manipulation. The trace depicts that the employee initially receives some inputs (e.g., indicated by the state property

    hears(employee, capable_colleagues_available)

at time point 1).

As a result, the employee has made some beliefs (e.g., the state property

    belief(employee, capable_colleagues_available)

at time point 2), which persists for a longer time. Next, when the employee receives a last minute request at time point 6

    hears(employee, last_minute_request)

he eventually generates desire to avoid the task at time point 8

    desire(employee, avoid_task)

Based on this desire and the input received earlier

    hears(employee, capable_colleagues_available)

the employee generates the intention to avoid the task at time point 9:

    intention(employee, avoid_task)

Based on this intention and the input received earlier

    hears(employee, own_agenda_full)

at time point 1, the employee eventually performs the action of avoiding the task at time point 10.

Figure 4 is the resulting simulation trace of the case study described in Section 5, in which the manager agent can reason and act in an anticipatory manner to avoid the employee's avoidance desire, intention and/or action to occur. Figure 4 shows that the manager initially desires that the employee does not perform the action to avoid the task:

    desire(manager, not(performs(employee, avoid_task)))

Based on this, he eventually generates number of more detailed desires about what the employee should hear (see, for example, the state property

    desire(manager, not(hears(employee, capable_colleagues_ available)))

at time point 3). Next, the manager uses these desires to generate some intentions to fulfil these desires (e.g., the state property

    intention(manager, tell(employee, not(capable_colleagues_ available)))

at time point 4). Eventually, these intentions are performed, and the employee receives some new inputs (e.g., the state property

    hears(employee, not(capable_colleagues_available))

at time point 7). As a result, when the employee receives a last minute request at time point 11

    hears(employee, last_minute_request)
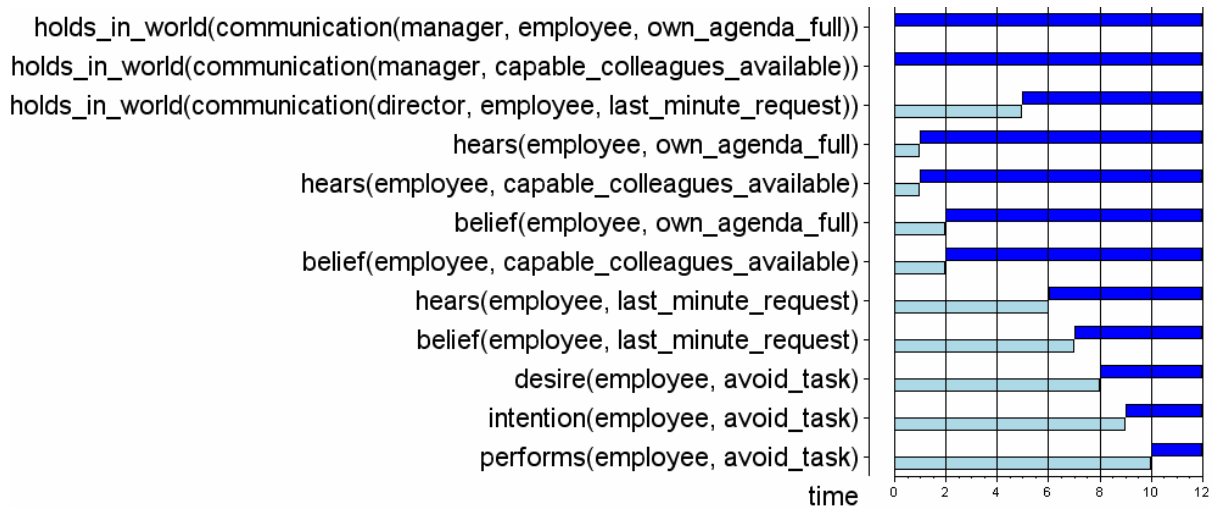
he does not generate the action to avoid the task.

Figure 3. Simulation trace of task avoidance behaviour (without ToM)

Figure 4. Simulation trace where task avoidance behaviour is prevented by social manipulation

Note that in the scenario sketched in Figure 4, the manager takes all possible actions (within the given conceptualisation) to fulfil its desires. This is a rather extreme case, since according to the employee's BDI-model, modifying only one of its input will be sufficient to make sure that (s)he does not avoid the task. Other traces can be generated in which the manager takes less actions to fulfil its desires.

7 RELATION TO OTHER WORK

The two-level BDI-Agent Model for Theory of Mind presented here is an example of recursive modeling and its use in social manipulation; see also [6], [8], [12], [17]. In the field of Agent Theory, the idea of recursive modeling has been described in [12] as follows:

'Recursive modeling method views a multiagent situation from the perspective of an agent that is individually trying to decide what physical and/or communicative actions it should take right now. […] In order to solve its own decision-making situation, the agent needs an idea of what the other agents are likely to do. The fact that other agents could also be modeling others, including the original agent, leads to a recursive nesting of models.' [12], p.125

In [17], PsychSim - an implemented multi-agent based simulation tool for modeling interactions and influence among groups or individuals - has been described in the context of childhood bullying and aggression, which provides interesting insight into the role that Theory of Mind plays in human behavior. In this work, an agent's theory of mind about other agents is crucial in the following sense:

'For example, a bully motivated by the approval of his classmates would use his mental model of them to predict whether they would enjoy his act of aggression and laugh along with him. Similarly the bully would use his mental model of the teacher to predict whether he will be punished or not' [17], p. 247

In PsychSim, agents maintain models of each other's beliefs, goals, policies, etc., and are able to reason about it. This is a form of recursive agent modeling specifically organised to model psychological factors that play a role in influencing human communication and human social interaction in general.

The two-level BDI-Agent model for Theory of mind for social manipulation presented in this paper can be considered a recursive model with two levels of nesting. At level 1, the manager uses BDI-concepts *within* the Theory of Mind to describe the reasoning process of the employee. At level 2, the manager uses BDI-concepts for reasoning *about* the Theory of Mind for its own reasoning (meta-reasoning) about the reasoning process of the employee.

The work by [8] considers many topics, like foundation of sociality (cooperation, competition, groups, organization, etc), levels of coordination and cooperation, emergent pre-cognitive structures and constraints. Specifically it addresses influencing other agents and trying to change their behavior based on Theory of Mind of the agent:

'The explicit representation of the agents mind in terms of beliefs, intentions, etc., allows for reasoning about them, and – even more importantly – it allows for the explicit influencing of others, trying to change their behavior (via changing their goals/beliefs).[…] The agents should have some decision function (that implicitly or explicitly presupposes some goal/desire/preference). The influencing agent should give them some hints for this decision, in order to change their behavior' [8], p. 178

However, in that work no formalisation is presented. In contrast, the model presented here has been formally specified.

8 DISCUSSION

In order to function efficiently in social life, it is very helpful for an agent to have capabilities to predict in which circumstances the agents in its environment will show certain behaviours. To this end, such an agent will have to perform reasoning based on a Theory of Mind [3]. This paper presents a model for reasoning based on a Theory of Mind, which makes use of BDI-concepts at two different levels. First, the model uses BDI-concepts *within* the Theory of Mind (i.e., it makes use of beliefs, desires and intentions to describe the reasoning process of another agent). Second, it uses BDI-concepts for reasoning *about* the Theory of Mind (i.e., it makes use of beliefs, desires and intentions to describe an agent's meta-reasoning about the reasoning process of another agent). At this second level, meta-statements are involved, such as 'B believes that A desires d' or 'B desires that A does not intend a'. These meta-statements are about the states occurring within the other agent. In addition, meta-statements are involved about the dynamics occurring within the other agents. An example of such a (more complex) meta-statement is 'B believes that, if A performs a, then earlier he or she intended a'.

The two-level BDI-based model as presented can be exploited both for *social anticipation* (i.e., in order to be prepared for the behaviour of another agent) and for *social manipulation* (i.e., in order to affect the behaviour of another agent at forehand). The model has been formalised using the high-level modelling language LEADSTO, which describes dynamics in terms of direct temporal dependencies between state properties in successive states. Based on the formal model, a number of simulation experiments have been performed within a specific case study, addressing the scenario of a manager that reasons about the task avoiding behaviour of his employee. A main difference with the earlier work described in [15] is that in the current paper the agent model is executable and therefore can easily be used for simulation. Moreover, it not only addresses reasoning about the other agent's beliefs, desires and intentions, but also integrates this with reasoning about the agent's own beliefs, desires and intentions, and actions in order to perform social manipulation. This part was not formalised in [15].

The case study illustrates how the two-level model can be used for social manipulation. For this purpose, the crucial steps are to find out which situations would lead to undesired behaviour of another agent, and to prevent these situations from occurring (or, similarly, to establish situations that would lead to desired behaviour). In addition, the model can be used for social anticipation. In that case, the main steps are to predict the behaviour that another agent will show given the current situation, and to prepare for this. Also this second type of reasoning based on a Theory of Mind is essential to function smoothly in social life.

In the model, not only Agent B has a Theory of Mind about Agent A, but it also uses this theory in its own reasoning process in order to do social manipulation, i.e., to change the behavior of Agent A. The model was designed in such a way that the Agent A does not know beforehand that Agent B is trying to manipulate him by changing some of the beliefs. Thus, the situation was not considered that Agent A tries to not to be manipulated by Agent B. In future research it will be addressed how such elements can be included in the model. For instance, the employee may have

models of other employees to infer who else is available for a task.

In [8], a basic problem of social life among cognitive agents has been addressed, that is: *how to induce the other agent to believe on us and even to do something*, or in other words: *why should he care about our goals and expectations?* This problem can be solved normally – but not necessarily – by communicating. Generally, in order to induce another agent to do or not to do something, we need power of influencing him. However, the most important basis of our power is the fact that also our actions are potentially interfering with his goals. This can be exploited to change his mind and induce him to do or not to do something. In the case study considered here, as the manager has power over his employee, he interfered in the goal of employee: by performing the action of changing his beliefs by communicating to him. As a result, the behavior of the employee has been changed, i.e., he doesn't avoid the task.

From a theoretical angle, much literature is available in foundations of approaches as the one presented here. For example in literature such as [13], [18], [19], a modal logic perspective is used to obtain formal semantics for languages that allow to express that an agent has reflective knowledge about what another agent knows. However, most of such modal logic approaches do not address the dynamics of the agents' processes in an executable manner. An exception is [4], where executable temporal logic is used as a basis; however, there the reflective aspect is not incorporated.

REFERENCES

[1] Aristotle (350 BC). *Nicomachean Ethics* (translated by W.D. Ross)

[2] Aristotle (350 BC). *De Motu Animalium* On the Motion of Animals (translated by A. S. L. Farquharson)

[3] Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

[4] Barringer, H., M. Fisher, D. Gabbay, R. Owens, & M. Reynolds (1996). *The Imperative Future: Principles of Executable Temporal Logic*, Research Studies Press Ltd. and John Wiley & Sons.

[5] Bogdan, R.J. (1997). *Interpreting Minds*. MIT Press.

[6] Boella, G. and van der Torre, L. (2004) Groups as agents with mental attitudes. In: Jennings, N.R., Sierra, C., Sonenberg, L., and Tambe, M. (eds.), *Proceedings of the third international joint conference on Autonomous Agents and Multi Agent Systems, AAMAS'04*, pp. 964-971.

[7] Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2007). A Language and Environment for Analysis of Dynamics by Simulation. *International Journal of Artificial Intelligence Tools*. To appear, 2007. Shorter, earlier version in: Eymann, T., Kluegl, F., Lamersdorf, W., Klusch, M., and Huhns, M.N. (eds.), *Proc. of the Third German Conference on Multi-Agent System Technologies, MATES'05*. LNAI, vol. 3550. Springer Verlag, 2005, pp. 165-178.

[8] Castelfranchi C. *Modelling Social Action for AI Agents*. Artificial Intelligence, vol. 103, 1998, pp. 157-182

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

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

[11] Georgeff, M. P., and Lansky, A. L. (1987). Reactive Reasoning and Planning. In: Forbus, K. and Shrobe, H. (eds.), *Proceedings of the Sixth National Conference on Artificial Intelligence*, AAAI'87. Menlo Park, California. American Association for Artificial Intelligence, 1987, pp. 677-682.

[12] Gmytrasiewicz P. J. and Durfee. E. H. A rigorous, operational formalization of recursive modeling. In: Lesser, V. (ed.), *Proceedings of the First International Conference on Multiagent Systems*, pp. 125-132, 1995.

[13] Halpern, J.Y., Fagin, R., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge*. MIT Press, 1995.

[14] Jonker, C.M., Treur, J., and Wijngaards, W.C.A., (2003). A Temporal Modelling Environment for Internally Grounded Beliefs, Desires and Intentions. *Cognitive Systems Research Journal*, vol. 4(3), 2003, pp. 191-210.

[15] Jonker, C.M., Treur, J., and Vries, W. de, (2002). Temporal Analysis of the Dynamics of Beliefs, Desires, and Intentions. *Cognitive Science Quarterly*, vol. 2, 2002, pp.471-494.

[16] Malle, B.F., Moses, L.J., Baldwin, D.A. (2001). *Intentions and Intentionality: Foundations of Social Cognition*. MIT Press.

[17] Marsella, S.C., Pynadath, D.V., and Read, S.J. *PsychSim: Agent-based modeling of social interaction and influence*. In: Lovett, M., Schunn, C.D., Lebiere, C., and Munro, P. (eds.), *Proceedings of the International Conference on Cognitive Modeling, ICCM 2004*, pp. 243-248 Pittsburg, Pensylvania, USA.

[18] Rao, A.S. and Georgeff, M.P. (1995) BDI-agents: from theory to practice. In: Lesser, V. (ed.), *Proceedings of the International Conference on Multiagent Systems*, pp. 312 – 319.

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