Modeling Agents with a Theory of Mind*

Maaike Harbers[1,2], Karel van den Bosch[2] and John-Jules Meyer[1]
[1]Utrecht University, P.O.Box 80.089, 3508 TB, Utrecht, The Netherlands
[2]TNO Human Factors, P.O.Box 23, 3769 ZG, Soesterberg, The Netherlands
{maaike,jj}@cs.uu.nl, karel.vandenbosch@tno.nl

Abstract

Training systems with intelligent virtual agents provide an effective means to train people for complex, dynamic tasks like crisis management or firefighting. Virtual agents provide more adequate behavior and explanations if they not only take their own goals and beliefs into account, but also the assumed knowledge and intentions of other players in the scenario. This paper describes a study to how agents can be equipped with a theory of mind, i.e. the capability to ascribe mental concepts to others. Based on existing theory of mind theories, a theory-theory (TT) and a simulation-theory (ST) approach for modeling agents with a theory of mind models are proposed. Both approaches have been implemented in a case study, and results show that the ST approach is preferred over the TT approach.

1 Introduction

Virtual training systems are often used to train people for complex, dynamic tasks in which fast decision making is required, e.g. commanding in crisis management, military missions or firefighting. In a training session, trainees are confronted with an incident or problem which they have to solve. To accomplish this task or mission, they have to interact with several virtual characters, e.g. colleagues, team-members or opponents. The roles of these characters are sometimes played by instructors or co-trainees, but in an increasing number of systems the characters' behavior is generated by intelligent agents. Using intelligent agents instead of humans increases training flexibility and may reduce personnel costs.

At the stage trainees start training with scenarios, they are expected to already have knowledge about the procedures in the domain, e.g. the division of tasks, and where to find which information. The challenge is to apply this knowledge in a realistic scenario. In this process, interaction with others plays an important role, especially when the players are dependent on each others' actions for achieving their own tasks, and thus requires believable behavior of the intelligent agents. Furthermore, the agents should also adapt their behavior to the trainee's performance to adjust the difficulty of the scenario to the trainee's skills.

When a virtual training system is used independently, thus not with an instructor, trainees should be supported in understanding the played scenario by the system. This can be accomplished by letting the virtual agents explain their actions. Several accounts of self-explaining agents for virtual training have been proposed, e.g. [22, 30, 14, 17]. After the training session is over, the agents can be queried or give explanations on their own initiative about the motivations behind their actions in the played session. The aim of such explanations is to give trainees better insight in the played training session.

Typical mistakes that occur during incident management include giving incomplete or unclear instructions, forgetting to monitor task execution, and failing to pick up new information and quickly adapt to it. Many of these errors involve situations in which people make false assumptions about others' knowledge or intentions. The tendency to attribute incorrect knowledge and intentions to others appears in stories of professionals [12], but it is also a well described phenomenon in general in cognitive sciences [23, 21].

In summary, virtual agents should be able to show believable behavior, adapt to the trainee's performance, give useful explanations, and thereby make trainees aware of the human tendency to attribute false mental concepts to others. In earlier work we have argued that these requirements can be met by equipping agents with a theory of mind [16]. Someone with a theory of mind has the ability to attribute mental states such as beliefs, intentions and desires to others in order to better understand, explain, predict or manipulate others' behavior. In this paper, we shortly discuss the uses of agents with a theory of mind in virtual training, but the focus of the paper is on their modeling and implementation. There are currently no agent programming

*This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

languages providing explicit constructs for the implementation of agents with a theory of mind. We will therefore propose and evaluate two approaches for modeling agents with a theory of mind, one based on theory-theory (TT) and another on simulation-theory (ST).

The outline of the paper is as follows. In section 2, we give an example of a training situation and explain in which ways agents with a theory of mind can enhance virtual training. The example serves as motivation for the use of agents with a theory of mind, but also as a specification of the criteria according to which the proposed agent models can be evaluated. In section 3, we give a short overview of theory of mind research and zoom in on two theories of theory of mind. In section 4, we translate the two theories to models of software agents, i.e. two approaches for implementing agents with a theory of mind. In section 5, we describe a case study in which the two approaches are evaluated. In section 6 and 7, we end the paper with a discussion and a conclusion, respectively.

2 A training scenario

The example in this section is a part of a virtual training scenario for on-board firefighting[1]. The trainee plays the role of H-Officer, the person in command when there is a fire aboard of a navy frigate. Besides the trainee, two others are involved, an E-Officer and an A-Officer, played by intelligent agents. The H-Officer leads the incident management from the Technical Center of the ship. His tasks involve assessing the situation, developing a plan, instructing other officers, monitoring task execution, and adapting plans if necessary. The E-Officer is also located at the Technical Center and is responsible for the electricity at different compartments of the ship. The A-Officer leads the fire attack at the location of the incident and can only use water in compartments where the electricity has been switched off. The H-Officer can communicate with all officers and vice versa, but there is no direct communication between the E-Officer and A-Officer possible.

In the optimal situation, if there is a fire, the E-Officer switches off the electricity in the right compartments and reports this in person to the H-Officer. Subsequently, the H-Officer broadcasts the message to the ship, and the A-Officer orders his team to attack the fire with water. As a result, the fire will be extinguished, which the A-Officer reports to the H-Officer. In this scenario course, the agents understood each others' and the trainee's goals, and acted proactively to support each other. The trainee received positive feedback in the form of a good end result, and explanations of the agents can even increase his understanding the played session. For instance, the E-Officer may explain that he switched off electricity to ensure that the A-Officer could safely attack the fire with water. By such explanations the trainee learns not only which but also why certain procedures have to be followed.

The scenario may also unfold otherwise, for example, when the trainee fails to broadcast the E-Officer's message. In such a case, it might be useful if the A-Officer asks the trainee whether the electricity has been switched off. The trainee will become aware of his failure and no longer delay the fire attack. A useful explanation for the A-Officer's action could be that it believed that the trainee would know about the status of the electricity.

For more advanced trainees, mistakes of virtual agents can create interesting learning situations. The E-Officer could for example fail to switch off electricity, forget to report to the trainee, or switch off electricity in a wrong compartment. The trainee is challenged to correct the agents, for instance by asking the E-Officer whether he already switched off electricity. An explanation of the E-Officer's failure could be that he believed that the A-Officer did not plan to use water for his fire attack.

Though the given situation is a simple one, several capabilities are required to provide training as described above. The intelligent agents should be able to attribute mental states to others, know when to help the trainee, make believable mistakes, and explain their own actions by their assumptions about other agents' states. In the example, interaction plays an important role and the different agents (including the trainee) are dependent on each other for successful task execution. In order to generate and explain the behaviors in the example, the agents have to be aware of the others' tasks and the consequences of their actions for others. In other words, the agents need some theory about the other agents' mental states: a theory of mind.

3 Theory of mind research

To understand the social world around them, people interpret others' and their own actions in terms of mental states. A theory of mind is the ability to understand others as intentional agents, and to interpret their minds in terms of intentional concepts such as beliefs and desires, e.g. *R believes that M intends him to persuade A that p*. The term 'theory of mind' originates from Premack and Woodruff's famous paper 'Does the chimpanzee have a theory of mind?' [26]. Since then, the term has been used to denote the research field in which the ability to explain and predict ones own and others' behavior is studied. Besides biologists, researchers from several other fields have been involved in theory of mind research, such as neuroscientists, psychologists and philosophers.

Humans are not born with a fully developed theory of

[1] The scenario is inspired on the Carim system, a virtual training system developed by TNO and VSTEP for the Netherlands Navy. For an overview of the system see [29].

mind, but acquire one during their childhood. The false-belief task [31] is often used by developmental psychologists to determine whether someone has a fully developed theory of mind. To test whether a child passes the task, an experimenter puts an object in a box in presence of the child and another person. The other person leaves the room and when she is gone, the experimenter puts the object in a different box. When the person returns the child is asked where she will look for the object. The child fails if it answers that the person will look in the second box. Though the child knows that the object is in the second box, to pass the task it should be able to understand that the other person did not see that the object was replaced and thus will look in the first box. Experiments demonstrated that children obtain the ability to perform this task well around the age of four years old.

Another contribution of psychology to theory of mind research are studies about the absence of a theory of mind, also called mind-blindness, with autists [1]. A mind-blind person has difficulties to determine the intentions of others and lacks understanding of how his behavior affects others.

Though psychologists studied theory of mind acquirement and theory of mind impairment, most of them remained neutral on how a fully developed theory of mind in adults works. Philosophers, in contrast, are focusing on exactly this question. Currently, the debate involves two prominent accounts on human, adult theory of mind: theory-theory and simulation-theory. According to theory-theorists (e.g. [8]), we have an implicit *theory* of the structure and functioning of the human mind. This theory involves a set of concepts, e.g. beliefs, desires and plans, and principles about how these concepts interact, e.g. people act to fulfill their desires. This theory allows us to understand, explain and predict our own, and other people's behavior. The mental states attributed to others are unobservable, but knowable by intuition or insight. Theory-theory relates to folk psychology, which refers to the way humans *think* that they reason [4]. Namely, humans use concepts such as beliefs, goals and intentions to understand and explain their own and others' behavior.

Simulation-theory (e.g. [13, 15]) was proposed as an alternative to theory-theory. According to simulation-theorists, theory of mind is the ability to project ourselves into another person's perspective, and simulate his or her mental activity with our own capacities for practical reasoning. Thus instead of a theory, theory of mind is a kind of knowledge that allows one to mimic the mental state of another person. In order to simulate another's mental processes, it is not necessary to categorize all the beliefs and desires attributed to that person as such. In other words, it is not necessary to be capable of complete introspection.

Whether human theory of mind follows the theory-theory or simulation-theory approach cannot be determined by just observing human adult behavior. Therefore, philosophers became interested in theory of mind development. According to some theory-theorists, acquiring a theory of mind is a matter of maturation of an innate module, which happens automatically. Others think it is instantiated through social interactions. According to simulation-theorists, the ability to simulate is innately given. Children only have to learn which of their mental states to vary when simulating, in order to adopt the right perspective.

There are several proposals for a mix of theory-theory and simulation-theory (e.g. [19, 24]). Many simulation-theorists argue for their position on grounds of simplicity. They claim that simulation is more efficient than acquiring a complete theory. For these reasons, some adherers of theory-theory admit that at least some form of simulation must take place when people reason about others, and incorporate simulation aspects into a theory-theoretic account. Though this makes theory-theory acceptable for some, others remain convinced that simulation forms the basic mechanism of theory of mind. Critics of simulation-theory however argue that in order to simulate, we must know what to simulate and for that a theory is needed. This resulted in approaches stating that others' behavior is predicted by simulation, but in addition, a body of theoretical knowledge is needed to govern these simulations.

4 Agents with a theory of mind

Inspired on the philosophical theories discussed, we propose two approaches for modeling agents with a theory of mind, one based on theory-theory and another on simulation-theory.

4.1 A theory-theory approach

Folk psychology, in which behavior of others is understood in notions like beliefs, desires and intentions, forms the basis of the theory-theory account of theory of mind. The theory-theory approach clearly relates to the BDI (belief desire intention) paradigm, which is used for modeling agents [28]. There is no single BDI model, but there are several agent programming languages based on the BDI paradigm, e.g. Jack [7], Jadex [25], Jason/AgentSpeak [2] and 2APL [10]. A typical BDI agent has a goal base, plan base, plan library and intentions, and those form the elements of its reasoning. The upper part of figure 1 shows the general architecture of a BDI agent.

The behavior of a BDI agent is directed by its goals. Dependent on its beliefs, the agent selects particular plans from its plan library to achieve theses goals. A plan is a recipe for achieving a goal given particular preconditions. The plan library may contain multiple plans for the achievement of one goal. An intention is the commitment of the agent to

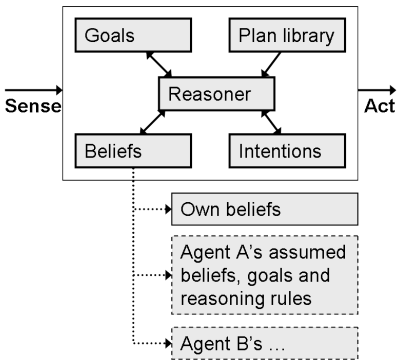

Figure 1. Theory-theory: architecture of a BDI agent with a theory of mind.

execute the sequence of steps making up the plan. A step can be an executable action, or a sub-goal for which a new plan should be selected from the plan library. A typical BDI execution cycle contains the following steps: i) observe the world and update the agent's internal beliefs and goals accordingly, ii) select applicable plans based on the current goals and beliefs, and add them to the intention stack, iii) select an intention and iv) perform the intention if it is an atomic action, or select a new plan if it is a sub-goal.

A theoretic approach to model an agent with a theory of mind is to add beliefs about other agents to a BDI agent's belief base. In figure 1 this is shown by the boxes below the general BDI architecture. Besides its own beliefs, the agent has beliefs that form its theory of mind (dashed boxes). The agent in figure 1 has a mind theory of agent A and B, incorporating the believed beliefs and goals of agent A and B. For instance, the belief *A(B(X))* would represent that the agent believes that agent A believes X, and *B(G(Y))* that the agent believes that agent B has goal Y. An agent's behavior is determined by its goals and beliefs. Thus, when an agent has beliefs about other agents, its behavior is also based on the believed beliefs or goals of others.

Besides beliefs about others' beliefs and goals, the agent must have a theory about how these elements interact. For instance, to predict someone's behavior, an agent needs to be able to make combinations of beliefs and goals, and derive new (sub-)goals, plans or actions. In this theory-based agent model, the rules according to which the elements combine are also added as beliefs to the agent's belief base. In other words, beliefs that make combinations between beliefs about another agent's beliefs and beliefs about that agent's goals are added. Such a reasoning rule belief is for example *if (A(B(X)) and A(G(Y))) then Z(P(Z))*, meaning that if one believes that agent A believes that B has goal Y, one can assume that agent A will execute plan Z. With these 'theory of mind beliefs', the agent is able to predict and explain other agents' behavior. To do so, the agent does not use its own practical reasoning power (the reasoner in figure 1), but instead, it uses its epistemic reasoning power for making inferences of its beliefs (the epistemic reasoner is part of the agent's belief base, and is not explicitly shown in figure 1).

4.2 A simulation-theory approach

The essence of simulation theory is that an agent uses its own reasoning power to reason about other agents, and thus not all of the other's reasoning steps have to be incorporated in a theory. Figure 2 shows a schematic picture of a theory of mind model based on the simulation-theory approach. Like in the theory-theory model, the agent has a reasoner which deliberates with the content of its mental states. In this picture however, the exact representation of mental states is not specified. Besides a representation of its own mental states, the agent has representations of mental states attributed to other agents (dashed boxes). The agent can take its own decision making system off-line, and start deliberating with the mental state of another agent to make predictions about its behavior.

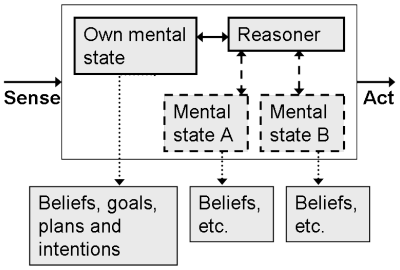

Figure 2. Simulation-theory: architecture of an agent with a module-based theory of mind.

Simulationists argue that in order to have a theory of mind, one does not need to have access to all reasoning rules according to which the other is reasoning. Radical simulationists even claim that the mental state of the other agent does not necessarily have to be organized in terms of beliefs and goals. Therefore, in figure 2 we have not specified how a mental state is represented, but in our approach we will assume that it is in terms of beliefs, desires and intentions, as shown in the boxes outside of the agent's internals.

The architecture in figure 2 can best be implemented in a module-based programming language. Each mental state, the agent's own and those of other agents, can be represented in a separate module. By using modules, the same practical reasoner can be used to reason with different mental states without interferences among them. If an agent wants to make a prediction about someone else's behavior, it just applies its reasoner to the assumed mental state of that agent. The agent thus reasons with another agent's mental concepts as if they are its own. The agent can use its as-

sumptions about the other agent as input for its own reasoning process and let its actions depend on them.

As we chose to specify mental states in terms of beliefs, goals and intentions in the simulation-based approach as well, a BDI-based agent programming language can also be used for the implementation. There are several BDI-based agent programming languages that allow for modularity, e.g. Jack [6], Jadex [5] and extended 2APL [11]. One of these could thus be used to implement the theory of mind model based on simulation-theory.

5 A case study

In section 2, we introduced a training example illustrating how agents with a theory of mind can improve virtual training. We sketched how the agents should behave in a specific training situation, e.g. performing support actions, making mistakes due to an incorrect theory of mind, and providing explanations based on a theory of mind. In section 4, we have introduced two approaches for modeling and implementing agents with a theory of mind. In this section, we describe a case study to evaluate whether the proposed approaches are indeed able to generate behavior and explanations such as described in section 2. The first part discusses the implementation of the agents, and the second part discusses the results of simulations with the agents.

5.1 Implementation

We used an extended version of the training scenario introduced in section 2, in which electricity needs to be switched off stepwise instead of in all compartments at once. We specified the desired behavior and explanations of the E-Officer and A-Officer for three variants on the scenario: optimal, support and challenge. In the optimal scenario nothing goes wrong, in the support scenario the trainee makes mistakes and the agents give support, and in the challenge scenario the agents make mistakes due to an incorrect theory of mind. Subsequently, we implemented three versions of the E-Officer and A-Officer agent: with no theory of mind (NT), based on the theory-theory approach (TT), and based on the simulation-theory approach (ST). We included the agent model with no theory of mind in the study to check whether equipping agents with a theory of mind has added value.

NT agents. We implemented the E-Officer and A-Officer agents with no theory of mind according to a methodology for developing self-explaining agents in virtual training [17] in the agent programming language 2APL [10]. The methodology describes how the tasks of the agents can be represented in a hierarchical structure, and how such a goal tree can be implemented in a BDI-based agent programming language. It ensures that an agent's reasoning steps are explicitly represented, and thus those reasoning steps responsible for generating an action can also be used to explain the same action. For example, an agent opens a door because it has the goal to save victims and it believes that there is a victim behind the door. Though the NT A-Officer and E-Officer agents could perform actions that had a positive effect on others' task execution, the agents' reasoning did not involve possible mental states of other agents. Information about other agents was thus implicitly present in the goal tree of the agents.

TT agents. We used the NT agents as a starting point for the implementation of the agents based on the theory-theory model, and extended them with beliefs about other agents' mental concepts and reasoning rules. The goal trees underlying the agents' implementations remained the same, but the conditions under which certain goals were adopted were changed. Namely, the conditions of goals which achievement had effect on other agents' task execution involved believed mental concepts about the others. For example, the E-Officer only switches off electricity in a compartment if it believes that someone else intends to use water there. The following 2APL code shows part of the E-Officer's theory of mind about the A-Officer in its belief base. In 2APL, an agent's belief base is a Prolog program.

```
a_off(g,extinguishFire).

a_off(b,noElectricityInComp37).
a_off(b,fireInComp38)).

a_off(p,attackWithWater) :-
    a_off(g,extinguishFire),
    a_off(b,noElectricityInComp37)),
    a_off(b,fireInComp38).
```

The first line of code represents a belief about a goal attributed to the A-Officer, the second and third beliefs are attributed beliefs, and the last belief incorporates a reasoning rule telling which plan the A-Officer will probably adopt when it has the corresponding beliefs and goal. The E-Officer can use its theory of mind for example by only switching off electricity in compartment 37, if the belief *a_off(p,attackWithWater)* is derivable from its belief base.

The TT agents have a first order theory of mind, which means that their theories of mind do not involve other agents' theories of mind. Thus, the agents have no beliefs like 'I believe that agent A believes that I have goal Y'. In this simple scenario, it was not necessary to implement agents with a second or higher order theory of mind, but there are no practical reasons against it.

ST agents. The E-Officer and A-Officer agents based on simulation-theory were implemented in Extended

2APL [11] instead of 2APL. In Extended 2APL, an agent can create modules, update modules with beliefs and goals, execute modules, and query their belief and goal bases. In our case, execution of a module might result in updating its belief, goal or intention base, but not executing actual actions in the environment. For instance, the following Extended 2APL code represents the E-Officer's plan for creating, updating and executing a module with a theory of mind of the A-Officer.

```
create(a_off, a_off);
a_off.updateBB(noElectricityInComp37);
a_off.execute(B(fire(Y)));
Update(noElectricity,fire(Y))
```

The first action creates an instantiation of the module a_off which also has the name a_off. The second action updates the instantiation with the belief *noElectricity*. Then, the module is executed till the stopping condition B(fire(Y)) is satisfied, meaning that the belief *fire(Y)* can be derived from the module's belief base. In the module a_off, the variable Y can have the values *burning* and *extinguished*. Finally, the result of the execution is updated to the agents own belief base, e.g. resulting in the belief *a_off(noElectricityInCompr37,fire(extinguished))*, which means that if the A-Officer believes that the electricity is switched off, the fire will be extinguished. The E-Officer agent could use its theory of mind when adoption of goals for switching off electricity depends on its beliefs with predictions about the A-Officer's behavior.

The ST agents also have a first order theory of mind. For the theory of mind modules we used the implementation of the NT agents. The ST A-Officer's theory of mind contained the NT E-Officer's mental states and vice versa. Also for ST agents holds that it is possible to implement agents with second or higher order theory of mind.

Trainee and environment. Finally, we implemented a trainee agent which could act as it should or make errors. The trainee agent had no theory of mind and could not give explanations. In general, most actions were communication actions, implemented in 2APL and Extended 2APL as send-message actions. Some of the actions were executed in the environment, e.g. switching off electricity. However, as such actions were rare, we did not connect the agents to an actual environment. Instead, the actions only influenced the belief bases of the agents. For example, the A-Officer's action to command its team to attack a fire with water added a belief *extinguishedFire* to its belief base. It was assumed that actions could not fail.

5.2 Experimental results

As mentioned before, we specified three versions of the scenario, optimal, support and challenge, containing the desired actions and explanations of the agents. We ran three simulations for each version, with NT, TT and ST agents. In the optimal version, the trainee agent making no mistakes interacted in three simulation runs with two NT, TT and ST E-Officer and A-Officer agents. In the support version, the trainee agent making mistakes interacted with two NT, TT and ST agents. To run the challenge version, we adapted the different implementations of the A-Officer and E-Officer agents so that they would make mistakes. The trainee not making mistakes interacted with the adapted versions of the NT, TT and ST agents. During each simulation run, the E-Officer and A-Officer's actions and explanations were logged, and these logs were compared to the scenarios specified beforehand such as shown in table 1.

Specified behavior	**Actual behavior**		
Actions	**NT**	**TT**	**ST**
1. E-Off switches off elect. comp 37	√	√	√
2. E-Off reports to H-Off	√	√	√
3. H-Off broadcasts message	√	-	-
4. A-Off enters comp 37	√	√	√
Etc.	...	...	...
Explanations	**NT**	**TT**	**ST**
1. A-Off will ext. fire with water	X	√	√
2. H-Off wants to be updated	X	√	√
4. No electricity in comp 37	√	√	√
Etc.	...	...	...

Table 1. Desired and actual behavior of the NT, TT and ST agents in the optimal scenario.

The left column of table 1 shows a part of the desired actions and explanations in the optimal scenario. The last three columns show whether the agents' actions and explanations did (√) or did not (X) match the specified ones. Explanations 1, 2 and 4 explain actions 1, 2 and 4, respectively. Action 3 is not explained because the H-Officer is played by the trainee, and the trainee agent does not explain. For all nine simulations we found that the agents' *actions* in the simulation matched the specifications for 100%. Thus, independent of whether the agents had a theory of mind and which theory of mind model, they were all able to display the specified actions, including support actions and making mistakes due to an incorrect theory of mind.

The *explanations* of the agents with a theory of mind, the TT and ST agents, matched all of the specified explanations. The agents were able to incorporate beliefs and goals of others in their explanations. The explanations of the agents without a theory of mind, the NT agents, did not always match the specified ones. The NT agents only gave explanations in terms of their own beliefs and goals. For

some actions these explanations matched the specified ones (e.g. explanation 4 in table 1), but they did not when the actions had consequences for other agents (e.g. explanation 1 and 2). Thus, agents with a theory of mind were able to explain the consequences of their actions for other agents, also for support actions and mistakes, and agents without a theory of mind were not.

We did no simulation runs in which we combined different models of the E-Officer and A-Officer agent, e.g. an NT and a ST model, but this would have given the same results.

6 Discussion

The case study showed that agents with a theory of mind (TT and ST) have advantages over agents without a theory of mind (NT). Though all three agent types generated equal behavior, the agents with a theory of mind were also able to give explanations involving other agents' assumed mental states, and the agents without a theory of mind were not. Concerning observable agent behavior (including explanations), we did not find a difference between the theory-based and the simulation-based approach, and there are no reasons to assume that the outcome would be different for other scenarios. However, in our evaluation we only considered the perspective of the end user, the trainee.

Before we consider other perspectives, we should remark that there is no common methodology for validating models representing human behavior. First, because not much attention has been paid to the validation of human behavior representation models and the field is still immature [18]. Moreover, there are different model types which each require their own validation [32]. Currently, most models are evaluated by their intended use, that is, from the perspective of the end user [9]. However, besides the perspective of the end user, human behavior representation models can also be viewed from a psychological and a developer's perspective.

The psychological perspective considers how well human behavior is represented. This perspective is not relevant here, as the agents in virtual training systems do not have to generate behavior that is as human as possible. The agents should behave human-like, but they may e.g. make more errors than an average human if that serves a learning goal. Moreover, as discussed in section 3, there is no agreement on how the human theory of mind works.

The developer's perspective concerns the effectiveness and efficiency of model creation. Effective and efficient model creation reduces development costs of virtual training systems, and is thus a relevant feature of an approach for modeling agents with a theory of mind. There are standard works for the assessment of software quality, e.g. the IEEE Standard 1061 [20], but these are not specialized for human behavior representation models. Therefore, instead of using a standard method, we discuss our experiences with the implementation of the agents in the case study.

A first finding concerns the reuse of code. When implementing the theory of mind of a TT agent, we had to translate a BDI representation of a mental state to a Prolog representation, and practical reasoning rules to epistemic reasoning rules. Namely, a TT agent's theory of mind is about a BDI agent, but represented only by beliefs. For the implementation of an ST agent, no such translation had to be made. Instead, existing code of one agent could be used to implement the theory of mind of another. Though the extra work of implementing TT agents compared to ST agents was not much in our case study, the advantage of reuse of code increases with more complex agent models. We thus may conclude that concerning the reuse of code, the ST approach is preferred over the TT approach.

A second observation involves the introduction of errors related to theory of mind use into the agent models. The introduction of single errors was comparably easy to implement in both agent models. However, in the TT approach errors could only be included individually, and in the ST approach it was possible to introduce some structural errors. A structural error is for example that an agent does not take its theory of mind about another agent into account at all, or that an agent bases its behavior on a theory of mind of the wrong agent. Also on this point, the ST approach is preferred over the TT approach.

A final advantage of the ST approach over the TT approach concerns the nature of the agent models. In the case study, all agent models were purely symbolic and BDI-based. The TT approach can only deal with BDI models, as all attributed mental concepts need to be represented in the agent's belief base. The ST approach, in contrast, can also be applied to agent models involving a mix between symbolic and numerical representations.

7 Conclusion

We have introduced two approaches for modeling agents with a theory of mind, based on the theory-theory and the simulation-theory of mind. We have performed a case study in which we compared agents with no theory of mind, a theory-theory of mind and a simulation-theory of mind in an actual training scenario. We found that all agent types were able to display the behavior we specified, but only the agents with a theory of mind were able to provide explanations in which others' mental states were involved. From the perspective of the end user, there are no differences between the two theory of mind approaches, but from a developer's perspective, the simulation-theory has several advantages over the theory-theory approach.

Existing accounts of agents with a theory of mind do not involve the simulation-theory principles as proposed in this paper. In most approaches, the agents reason *about*

attributed mental concepts, and not *with* attributed mental concepts as if it were their own, e.g. [3, 27]. In future work, we will continue modeling agents with a theory of mind according to the simulation-based approach, and make them more complex, apply them to more diverse scenarios, and validate their use with human in the loop experiments. With these experiments we hope to demonstrate that agents with a theory of mind can contribute to trainees' learning performances.

References

[1] S. Baron-Cohen. *Mindblindness: an essay on autism and theory of mind*. MIT Press, Cambridge, 1995.

[2] R. Bordini, J. Hubner, and M. Wooldridge. *Programming multi-agent systems in AgentSpeak using Jason*. Wiley, 2007.

[3] T. Bosse, Z. Memon, and J. Treur. A two-level bdi-agent model for theory of mind and its use in social manipulation. In *Proceedings of the AISB 2007 Workshop on Mindful Environments*, pages 335–342, 2007.

[4] M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, Massachusets, 1987.

[5] L. Braubach, A. Pokahr, and W. Lamersdorf. Extending the capability concept for flexible bdi agent modularization. In *Proc. of ProMAS 2005*, pages 139–155, 2005.

[6] P. Busetta, N. Howden, R. Ronnquist, and A. Hodgson. Structuring bdi agents in functional clusters. In N. Jennings and Y. Lesperance, editors, *Intelligent Agents VI: Theories, Architectures and Languages*, pages 277–289, 2000.

[7] P. Busetta, R. Ronnquist, A. Hodgson, and A. Lucas. Jack intelligent agents - components for intelligent agents in java. *AgentLink News Letter*, 1999.

[8] P. Carruthers. *Theories of theories of mind*, chapter Simulation and self-knowledge: a defence of the theory-theory. Cambridge University Press, Cambridge, 1996.

[9] B. Chrandrasekaran and J. Josephoson. Cognitive modeling for simulation goals: A research strategy for computer-generated forces. In *Proc. of the 8th Computer Generated Forces and Behavioural Representation Conf.*, pages 239–250, 1999.

[10] M. Dastani. 2APL: a practical agent programming language. *Autonomous Agents and Multi-agent Systems*, 16(3):214–248, 2008.

[11] M. Dastani, C. Mol, and B. Steunebrink. Modularity in agent programming languages: An illustration in extended 2apl. Technical Report, 2008.

[12] R. Flin and K. Arbuthnot, editors. *Incident command: Tales from the hot seat*. Ashgate Publising, 2002.

[13] A. Goldman. In defence of the simulation theory. *Mind and Language*, 7:104–119, 1992.

[14] D. Gomboc, S. Solomon, M. G. Core, H. C. Lane, and M. van Lent. Design recommendations to support automated explanation and tutoring. In *Proc. of the 14th Conf. on Behavior Representation in Modeling and Simulation*, Universal City, CA., 2005.

[15] R. Gordon. *Theories of theories of mind*, chapter 'Radical' simulationism. Cambridge University Press, 1996.

[16] M. Harbers, K. Van den Bosch, and J. Meyer. Enhancing training by using agents with a theory of mind. In M. Beer, M. Fasli, and D. Richards, editors, *Proceedings of EduMas 2009*, pages 23–30, Budapest, Hungary, 2009.

[17] M. Harbers, K. Van den Bosch, and J. Meyer. A methodology for developing self-explaining agents for virtual training. In Decker, Sichman, Sierra, and Castelfranchi, editors, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 1129–1130, Budapest, Hungary, 2009.

[18] S. Harmon, D. Hoffmann, A. Gonzalez, R. Knauf, and V. Barr. Validation of human behavior representation. In *Proceedings of Workshop on Foundations for Modeling and Simulation (MS), Verification and Validation (VV) in the 21st Century*, Maryland, USA, 2002. Laurel.

[19] J. Heal. *Theories of theories of mind*, chapter Simulation, theory, and content. Cambridge University Press, 1996.

[20] IEEE. Standard for a software quality metrics methodology. *IEEE Std 1061-1998*, 1998.

[21] B. Keysar, S. Lin, and D. Barr. Limits on theory of mind use in adults. *Cognition*, 89:25–41, 2003.

[22] W. Lewis Johnson. Agents that learn to explain themselves. In *Proc. of the 12th Nat. Conf. on Artificial Intelligence*, pages 1257–1263, 1994.

[23] S. Nickerson. How we know -and sometimes misjudge- what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125(6):737–759, 1999.

[24] J. Perner. *Theories of theories of mind*, chapter simulation as explicitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix. Cambridge University Press, Cambridge, 1996.

[25] A. Pokahr, L. Braubach, and W. Lamersdorf. Jadex: A bdi reasoning engine. In R. Bordini, M. Dastani, J. Dix, and A. Seghrouchni, editors, *Multi-Agent Programming*. Kluwer Book, 2005.

[26] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1:515–526, 1978.

[27] D. Pynadath and S. Marsella. Psychsim: Modeling theory of mind with decision-theoretic agents. In *Proc. of the Internat. Joint Conf. on Artificial Intelligence*, pages 1181–1186, 2005.

[28] A. Rao and M. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proc. of the 2nd Internat. Conf. on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann publishers Inc., 1991.

[29] K. Van den Bosch, M. Harbers, A. Heuvelink, and W. Van Doesburg. Intelligent agents for training on-board fire fighting. To appear, 2009.

[30] M. Van Lent, W. Fisher, and M. Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proc. of IAAA 2004*, Menlo Park, CA, 2004. AAAI Press.

[31] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128, 1983.

[32] M. Young. Human performance model validation: One size does not fit all. In *Proc. of the Summer Computer Simulation Conf.*, pages 732–736, 2003.