Modeling agents with a theory of mind: theory-theory versus simulation theory

Maaike Harbers [a,b,*], Karel Van den Bosch [b] and John-Jules Meyer [a]

[a] *Information and Computing Sciences, Utrecht University, P.O.Box 80.089, 3508 TB, Utrecht, The Netherlands*
E-mail: {maaike,jj}@cs.uu.nl

[b] *Training Innovations, TNO Human Factors, P.O.Box 23, 3769 ZG, Soesterberg, The Netherlands*
E-mail: karel.vandenbosch@tno.nl

Abstract. Virtual training systems with intelligent agents provide an effective means to train people for complex, dynamic tasks like crisis management or firefighting. For successful training, intelligent virtual agents should be able to show believable behavior, adapt their behavior to the trainee's performance and give useful explanations about their behavior. Agents can provide more believable behavior and explanations if they, besides their own, take the assumed knowledge and intentions of other players in the scenario into account. This paper proposes two ways to model agents with a theory of mind, i.e. equip them with the ability to ascribe mental concepts such as knowledge and intentions to others. The first theory of mind model is based on theory-theory (TT) and the second on simulation theory (ST). In a simulation study, agents with no theory of mind, a TT-based theory of mind, and an ST-based theory of mind are compared. The results show that agents with a theory of mind are preferred over agents with no theory of mind, and that, regarding agent development, the ST model has advantages over the TT model.

Keywords: Theory of mind, Theory-theory, Simulation theory, BDI agents, Virtual training

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, 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 because that increases training flexibility and reduces personnel costs. As virtual training prepares trainees for situations in the real world, intelligent agents should display believable, realistic behavior [26].

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 [16], but it is also a well described phenomenon in general in cognitive sciences [31,28]. Thus, in order to display believable behavior, intelligent agents should have the ability to realistically (fail to) take others' assumed knowledge and intentions into account. This is especially important in teamwork [44], i.e. when players are dependent on each others' actions for achieving their own tasks.

When 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. In the beginning, it will be challenging for them to apply this knowledge

[*] Corresponding author: phone +31 (30) 253 2814, fax +31 (30) 253 4619.

in a realistic scenario in which all agents act as they should. Agents may even help the trainee when he fails to undertake required actions, e.g. by giving advices. At a later stage, when the trainee can easily play such scenarios, they can be made more challenging. Agents can start making mistakes, for instance, by attributing incorrect knowledge and intentions to others. Agents should thus be able to 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 the reasons for their actions. Several accounts of self-explaining agents for virtual training have been proposed, e.g. the Debrief explanation component [30], the XAI explanation components [41,19], and an approach by the authors of this paper [21,22]. After the training session is over, such agents can be queried or give explanations on their own initiative about the motivations behind their actions in the played session. To increase the trainee's understanding, the explanations should also include assumptions about others' mental states.

In summary, virtual agents should be able to show believable behavior, make trainees aware of the human tendency to attribute false mental concepts to others, adapt to the trainee's performance, and give useful explanations about their behavior. In previous work, the authors argued that these requirements can be met by equipping agents with a theory of mind [23]. 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, two models for agents with a theory of mind, one based on theory-theory (TT) and another on simulation theory (ST), will be proposed and evaluated.

The outline of the paper is as follows. Section 2 gives an example of a training situation and explains in which ways agents with a theory of mind can enhance virtual training. The example serves as a 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 will be evaluated. Section 3 gives an overview of theory of mind research and zooms in on two theories of theory of mind in particular: theory-theory and simulation theory. In section 4, two ways to model agents with a theory of mind are proposed, based on the two theories. Section 5 describes a case study in which agents with no theory of mind, a TT-based theory of mind, and an ST-based theory of mind are compared. Section 6 discusses related work and section 7 ends the paper with a conclusion.

2. An example training scenario

The present example is 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 other players are involved, an A-Officer and an E-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 E-Officer advices the trainee to broadcast the message or

[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 [40].

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. Background: theory of mind

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?' [35]. Since then, the term has been used to denote the research field in which the ability to explain and predict one's 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 mind, but acquire one during their childhood. The false-belief task [42] 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 [2]. 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 the question 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. [11]), people 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 [6]. Namely, humans use concepts such as beliefs, goals and intentions to understand and explain their own and others' behavior.

Simulation theory (e.g. [18,20]) 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 and took different views on it [10]. 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. [25,32]). Simulation theory is defended on grounds of simplicity. According to simulation-theorists, 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, it must be known 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. Two ways to model agents with a theory of mind

This section presents a theory-theory and a simulation theory approach for modeling agents with a theory of mind. The implementation of both approaches is also discussed, as there are currently no agent programming languages providing explicit constructs for the implementation of agents with a theory of mind.

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 intelligent agents [37]. There is no single BDI model, but there are several agent programming languages based on the BDI paradigm, e.g. Jack [9], Jadex [34], Jason/AgentSpeak [4] and 2APL [13]. 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.

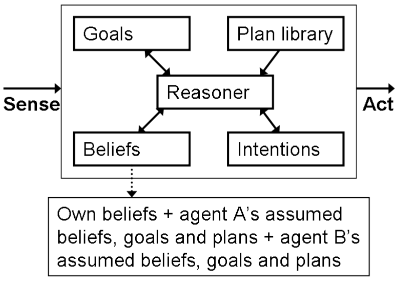
Goals
Plan library
Sense
Reasoner
Act
Beliefs
Intentions
Own beliefs + agent A's assumed beliefs, goals and plans + agent B's assumed beliefs, goals and plans

Fig. 1. Architecture of a BDI agent with a theory of mind based on theory-theory.

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 these 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 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.

An approach to model an agent with a theory of mind based on theory-theory 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 may have beliefs with mental concepts attributed to other agents (dashed boxes). The agent in Figure 1 has beliefs about attributed beliefs and goals of agent A and B. For instance, the belief *A(B(X))* represents 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 and 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 believed beliefs and goals, and derive new believed (sub-)goals, plans or actions. In this theory-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 A(P(Z))*, meaning that if the agent believes that agent A believes X and has goal Y, one can assume that agent A will execute plan Z. With these 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 simulation theory. Like in the theory-theory model, the agent has a reasoner which deliberates with the content of its mental state. Besides a representation of its own mental state, the agent has representations of mental states attributed to other agents (dashed boxes). The agent can take its own decision making system offline, and start deliberating with the mental state of another agent to make predictions about its behavior. In other words, it applies its own reasoner to the attributed mental states.

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 [20]. Therefore, in Figure 2 it is not specified how a mental state is represented, but in the present approach it is assumed that it is in terms of beliefs, desires and intentions, as shown in the boxes outside of the agent's internals.

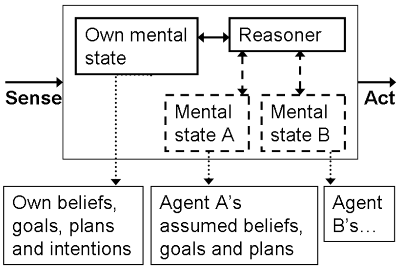
Own mental state
Reasoner
Sense
Act
Mental state A
Mental state B
Own beliefs, goals, plans and intentions
Agent A's assumed beliefs, goals and plans
Agent B's...

Fig. 2. Architecture of an agent with a theory of mind based on simulation theory.

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

As the agent's mental states are specified 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 [8], Jadex [7] and extended 2APL [14]. One of these could thus be used to implement the theory of mind model based on simulation theory.

4.3. Discussion

The two approaches just presented can both be implemented in a BDI-based agent programming language. For the TT approach, any BDI-based language can be used, and for the ST approach only those that allow for modularity. The most important difference between the TT and ST approach is the way to reason with attributed beliefs and goals. In the TT approach an epistemic reasoner is used, and in the ST approach a practical reasoner.

An advantage of the ST approach is that a practical reasoner can reason with attributed beliefs, goals and reasoning rules immediately, and the same mental state representations can be (re)used as 'normal' mental states and attributed mental states. In the TT approach, in contrast, 'normal' mental state representations have to be transformed to a different representation in order to reason with them as attributed mental states, which provides extra work for the programmer.

On the other hand, a reasoning process on an attributed mental state should have a result that the agent can use in it's own reasoning process, e.g. a belief in its belief base. In the TT approach, the attributed mental states are already represented by beliefs and no extra action is needed. In the ST approach, however, an extra action is needed to update the agent's own belief based with the result of a reason process on an attributed mental state. This makes the agent program more complex.

5. Comparing the two approaches

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 [24]. Moreover, there are different model types which each require their own validation [45]. Currently, most models are evaluated by their intended use, that is, from the perspective of the end user [12]. 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 the generation of human behavior in agents matches the generation of actual human behavior. The developer's perspective concerns the effectiveness and efficiency of model creation.

This section describes a study which compares the observable behavior of agents with no theory of mind, a theory of mind based on theory-theory, and a theory of mind based on simulation theory. It will be examined whether the different agents are able to generate the behavior and explanations required for the user. Thus, the user perspective is considered.

In the discussion, the developer perspective is taken into account as well. There are standard works for the assessment of software quality, e.g. the IEEE Standard 1061 [27], but these are not specialized for human behavior representation models. Therefore, instead of using a standard method, the authors' experiences with the implementation of the agents in the case study will be presented in the discussion.

The psychological perspective will not be considered, 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.

5.1. Methods

To compare agents with different theory of mind models, training scenarios were used to specify which behavior the agents should perform. For the study, three variants of the training scenario in section 2 were specified: an optimal, a supporting and a challenging version. 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. Besides the agents' actions in the scenario, also the corresponding explanations were specified. The different scenarios will be described in more detail in section 5.1.1.

All three scenarios involved three agents: an A-Officer, an E-Officer and a trainee (playing the H-Officer). Three versions of the A-Officer and E-Officer agents were implemented: agents with no theory of mind (NT), agents with a theory-theory of mind (TT), and agents with a simulation theory of mind (ST). The agents were implemented such that they would generate the actions and explanations in the specified scenarios as much as possible. The implementation of the agents will be discussed in section 5.1.2.

After specifying the scenarios and implementing the agents, different simulation sessions with the agents were run. An overview of the different simulation runs is provided in section 5.1.3. In the simulations, for an agent to perform well, its actions and explanations in the simulations should match the actions and explanations specified beforehand. ting the agents, different simulation sessions with the agents were run. An overview of the different simulation runs is provided in section 5.1.3. In the simulations, for an agent to perform well, its actions and explanations in the simulations should match the actions and explanations specified beforehand.

5.1.1. Scenario specification

Table 1, 2 and 3 show the specification of the events and agents' actions and explanations in the optimal, support and challenge scenario, respectively. In the tables, A, E and H refer to A-Officer, E-Officer and H-Officer, mes(ne) stands for the message there is no electricity in compartment 37, mes(e) stands for the message there is electricity in compartment 37, and mes(fe) stands for the message the fire in compartment 37 is extinguished.

Actions / ***events***	**Explanations**
Alarm: fire in comp 37	
E switches off elect. comp 37	then A can ext. fire with water
E reports mes(ne) to H	then H can broadcast mes(ne)
H broadcasts mes(ne)	-
A enters comp 37	to ext. the fire in comp 37
A ext. fire with water	no electricity in comp 37
Fire extinguished	
A reports mes(fe) to H	then H can broadcast mes(fe)
H broadcasts mes(fe)	-

Table 1

Actions, events and explanations in the optimal scenario.

In Table 1, 2 and 3, the left column shows the actions and events of that scenario in chronological order, and the right column shows the desired explanations for actions of the A-Officer and E-Officer. In the first scenario, displayed in Table 1, none of the agents makes a mistake.

Actions / ***events***	**Explanations**
Alarm: fire in comp 37	
E switches off elect. comp 37	then A can ext. fire with water
E reports mes(ne) to H	then H can broadcast mes(ne)
Nothing happens	
E advices H: broadcast mes(ne)	then A can ext. fire with water
H broadcasts mes(ne)	-
A enters comp 37	to ext. the fire in comp 37
A ext. fire with water	no electricity in comp 37
Fire extinguished	
A reports mes(fe) to H	then H can broadcast mes(fe)
Nothing happens	
A advices H: broadcast mes(fe)	then crew will be informed
H broadcasts mes(fe)	-

Table 2

Actions, events and explanations in the support scenario.

In the second scenario, the H-Officer, that is to be played by the trainee agent, forgets to broadcast the message that the electricity has been switched off in compartment 37. The E-Officer supports the H-officer by advising him to do so. Later, the H-Officer again forgets to broadcast a message. Then the A-Officer advices him to inform the crew that the fire has been extinguished.

Actions / ***events***	**Explanations**
Alarm: fire in comp 37	
Nothing happens	
H asks E about elect. comp 37	-
E reports mes(e) to H	H asked about elect. comp 37
H orders E: switch off elect.	-
E switches off elect. comp 37	H ordered to switch off elect.
E reports mes(ne) to H	then H can broadcast mes(ne)
H broadcasts mes(ne)	-
A enters comp 37	to ext. the fire in comp 37
A ext. fire with water	no electricity in comp 37
Fire extinguished	
Nothing happens	
H asks A about status fire	-
A reports mes(fe) to H	H asked about status fire
H broadcasts mes(fe)	-

Table 3

Actions, events and explanations in the challenge scenario.

In the third scenario, the E-Officer and the A-Officer both make one error. The E-Officer does not switch of electricity in compartment 37 by itself, the H-Officer explicitly has to order him to do so. The A-Officer forgets to update the H-Officer when the fire has been extinguished.

5.1.2. Agent implementation

To implement the agents, an approach for developing self-explaining agents in virtual training described in previous work [21] was used. In the approach, the goals and tasks of an agent are represented in a hierarchical goal structure. Beliefs are added to the goal hierarchy to denote under which conditions goals are adopted and dropped. The approach describes how such goal hierarchies can be implemented in a BDI-based agent programming language. This approach makes agents self-explainable, as the beliefs and goals that were responsible for generating an action can also be used to explain that 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.

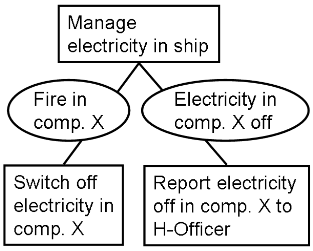
Manage electricity in ship
Fire in comp. X
Electricity in comp. X off
Switch off electricity in comp. X
Report electricity off in comp. X to H-Officer

Fig. 3. E-Officer agent without a theory of mind.

NT agents. Figure 3 shows a goal hierarchy of the E-Officer agent without a theory of mind. The boxes represent the agent's goals and the ovals represent its beliefs. The E-Officer's main goal is to manage the electricity in the ship, which is divided in the subgoals to switch off electricity and to report to the H-Officer that the electricity has been switched off. The beliefs in the hierarchy denote when the subgoals become active.

A similar goal hierarchy of the A-Officer with no theory of mind was made. Both NT agents were implemented in the BDI-based agent programming language 2APL [13]. In 2APL, an agent's mental state is defined by its beliefs, goal, plans and reasoning rules. Reasoning rules are generally of the form `Goal|Belief<-Plan`. The goal and belief(s) specified in `Goal` and `Belief` are checked against the agent's goal and belief base, respectively. When both return true (the agent has those goals and beliefs), the agent will execute the actions specified in `Plan`. Initially, the NT E-Officer agent has the following mental state.

```
Goals:
  manageE
Reasoning rules:
  manageE | comp(X,fire) <- switchOffE(X)
  manageE | comp(X,noE) <- reportToH(noE)
```

At the start of the scenario, the agent has one goal, managing the electricity, two reasoning rules, and no plans or beliefs. Only once the agent obtains the belief that there is a fire in compartment X or that it switched off the electricity in compartment X, it will generate plans.

Though the NT agents can perform actions that have a positive effect on others' task execution, the agents' reasoning does not involve possible mental states of other agents. Information about other agents is thus implicitly present in the NT agents' mental states.

TT agents. To develop agents with a theory of mind (both TT and ST), the NT agents were extended with a theory of mind ability. Figure 5.1.2 shows that a theory of mind ability, theory-theory-based or simulation theory-based, delivers extra beliefs due to theory of mind reasoning. By that, the adoption conditions for subgoals change. Namely, the conditions of goals which achievement have effect on other agents' task execution, involve believed mental concepts about the others. In the example, the E-Officer only switches off electricity in a compartment if it believes that someone else intends to use water to extinguish a fire in that compartment.

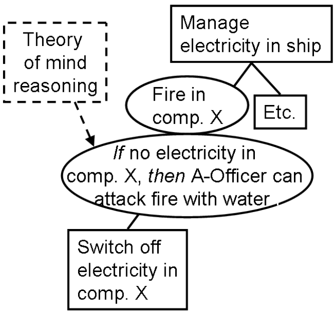
Theory of mind reasoning
Manage electricity in ship
Fire in comp. X
Etc.
If no electricity in comp. X, then A-Officer can attack fire with water
Switch off electricity in comp. X

Fig. 4. E-Officer agent with a theory of mind.

The way in which theory of mind reasoning is implemented, differs for TT and ST agents. TT agents' theory of mind is implemented in their belief base. The following 2APL code shows part of the E-Officer's theory of mind about the A-Officer in its belief base. Note that in 2APL, an agent's belief base is a Prolog program.

```
Beliefs:
  a_off(g,extinguishFire).

  a_off(b,comp(X,noE)).
  a_off(b,comp(X,fire)).

  a_off(p,attackFire(X,water):-
     a_off(g,extinguishFire),
     a_off(b,comp(X,noE)),
     a_off(b,comp(X,fire)).
```

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 these beliefs and goal. In other words, the E-Officer believes that the A-Officer will attack a fire in compartment X with water, when the electricity in that compartment has been switched off.

The E-Officer uses its theory of mind ability when predictions about the A-Officer's behavior will influence its own choices. This can be accomplished by adding extra belief checks its reasoning rules. The E-Officer's first reasoning rule then becomes as follows.

```
Reasoning rules:
  manageE | (comp(X,fire) and
    a_off(p,attackFireWithWater)) <-
    switchOffE(X)
```

The belief `a_off(p,attackFireWithWater))` is added to the belief check of the reasoning rule.

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 scenario, it was not necessary to implement agents with a second or higher order theory of mind, but it is practically possible.

ST agents. Like TT agents, ST agents are based on NT agents, but extended with a theory of mind (see figure). However, instead of TT agents, the theory of mind ability of ST agents is implemented by using modules. Therefore, the A-Officer and E-Officer agent based on simulation theory were implemented in Extended 2APL [14]. In Extended 2APL, an agent can create modules, update modules with beliefs and goals, execute modules, and query their belief and goal bases. For agents with a theory of mind, execution of a module might result in updating its belief, goal or intention base, but not in 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.

```
Plans:
   create(a_off, a_off);
   a_off.updateBB(comp(X,noE));
   a_off.execute(B(planAoff(Y)));
   Update(comp(X,noE),planAoff(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 `comp(X,noE)`. Then, the module a_off is executed till the stopping condition `B(planAoff(Y))` is satisfied, i.e. the belief `planAoff(Y)` can be derived from the module's belief base. The variable Y can have different values, representing a prediction of what the A-Officer's will do. During this execution, the execution of the agent owning the module, the E-Officer, is paused. In the last line of code, the result of the execution is updated to the agents own belief base, e.g. resulting in the belief `a_off(comp(X,noE), planAoff(attackFireWithWater))`, which means that if the A-Officer believes that the electricity is switched off, he will attack the fire with water.

Similar to TT agents, the E-Officer agent uses its theory of mind when the adoption of goals for switching off electricity depend on beliefs with predictions about the A-Officer's behavior.

Like the TT agents, the ST agents also have a first order theory of mind. For the theory of mind modules the implementation of the NT agents were used. 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.

5.1.3. Simulation runs

Several simulations were run to test whether the implemented agents were able to generate the actions and explanations specified in the three scenarios. Each agent type (NT, TT and ST) played each scenario (optimal, support and challenge) once, so in total nine (3x3) simulations were run. The A-Officer and E-Officer were always of the same type in one simulation run (both NT, both TT or both ST) because a combination of different agent types (e.g. NT and ST) would not have influenced the results. To run the challenging scenario, the implementations of the A-Officer and E-Officer agents were adapted so that they would make mistakes.

All characters in the scenarios were played by agents, and there were no humans involved in the simulations. Therefore, it was not needed to create a visualization of the agents and their environment. However, to simulate the three scenarios, besides the A-Officer and E-Officer, a trainee and an environment were needed.

Two versions of the trainee agent were implemented in 2APL. A trainee agent making mistakes was used for the support scenario, and one not making mistakes was used to run the optimal and the challenge scenario. The trainee agent consisted of a few simple rules reasoning rules, had no theory of mind and could not give explanations.

The role of the environment was minimized in the scenarios. The only two events in the environment in all three scenarios are a fire alarm and the extinction of the fire. Therefore, instead of implementing a separate environment, the events were represented in the belief bases of the agents. All agents believed that there was a fire in compartment 37 at the beginning of each simulation run. And the A-Officer's action to command its team to attack a fire with water led to addition of the belief *extinguishedFire* to its belief base. It was thus assumed that actions could not fail.

During the simulations, the actions and explanations of the agents were logged. The next section presents a comparison between the agents' actual behavior and their required behavior.

5.2. Results

During each simulation run, A-Officer and E-Officer's actions and explanations were logged, and these logs were compared to the specified scenarios. Table 4 shows the results of three of the three simulation runs of the optimal scenario, with the NT, TT and ST agents.

Specified behavior	**Actual behavior**		
Actions	**NT**	**TT**	**ST**
E switches off elect. comp 37	√	√	√
E reports mes(e) to H	√	√	√
A enters comp 37	√	√	√
A ext. fire with water	√	√	√
A reports mes(f) to H	√	√	√
Explanations	**NT**	**TT**	**ST**
then A can ext. fire with water	X	√	√
then H can broadcast message(e)	X	√	√
to ext. the fire in comp 37	√	√	√
no electricity in comp 37	√	√	√
then H can broadcast message(f)	X	√	√

Table 4

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

The left column of Table 4 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. The table only shows actions and explanations of the A-Officer and E-Officer, as those were the agents to be evaluated. Events and actions of the H-Officer are not displayed.

Table 4 shows that all of the agents' actions matched the specified ones. The simulation runs of the support and challenge scenarios had similar results. In all nine simulations it was 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 results in table 4 also show that 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. the third and fourth explanation in Table 4), but they did not when the actions had consequences for other agents (e.g. the first two explanations). 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. The same results were found in the support and challenge scenario.

5.3. Discussion

The results show 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), thus for the user, there was no 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.

For a developer, however, there are differences between TT and ST agents. A first observation concerns the reuse of code. When implementing the theory of mind of a TT agent, the BDI representation of a mental state had to be translated 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. It may thus be concluded that concerning the reuse of code, the ST approach is preferred over the TT approach.

A second finding 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.

6. Related work

An agent with a theory of mind has the ability to form models about other agents' mental states. These models can in turn contain models of other agents, which can contain models of other agents, etc. This is a form of recursive modeling, which was first introduced by Gmytrasiewicz and Durfee [17].

The theory of mind ability discussed in this paper focuses on forming and using attributions of beliefs, goals and plans of single agents. There are several executable theory of mind models which capture other aspects. For instance, Scasselati described an account of theory of mind for robots which focuses on gaze behavior [38], Peters introduced a theory of mind approach for conversation initiation in virtual environments [33], and Boella and Van der Torre presented an approach involving the attribution of mental attitudes to groups instead of single agents [3].

Bosse et al introduced a formal BDI-based agent model for theory of mind, and showed its use in modeling social manipulation, animal cognition and virtual character behavior [5]. Like the approaches presented in this paper, Bosse et al represent both the mental state of the attributing agent and its mental state attributions in terms of beliefs, desires and intentions. In Bosse et al's approach, agents reason *about* attributed mental concepts (in contrast to reasoning *with* attributed mental concepts as if it are one's own). This corresponds to the theory-theory approach described in this paper. Agents in Bosse et al's approach do not use their own reasoning power for reasoning with attributed mental concepts, like the agents based on simulation theory as presented in this paper.

PsychSim is a simulation tool for modeling interaction between agents with a theory of mind [36]. PsychSim agents have a decision-theoretic world model, including beliefs about their environment and recursive models of other agents. Where the models presented in this paper are based on a BDI model, the PsychSim agents are based on quantitative models of uncertainty and preferences. The PsychSim approach thus involves less symbolic representations of agents' mental states and is therefore less appropriate for explanation purposes than the BDI-based approaches presented here.

Laird introduced an account of theory of mind with the goal to add anticipation to a Quakebot [29]. In Laird's approach, an agent creates an internal representation of what it thinks the enemy's internal state is, based on its observation of the enemy. The agent then predicts the enemy's behavior by using its own knowledge of tactics to select what it would do if it were the enemy. As in the simulation theoretic approach presented in this paper, the agent reasons as if it were the other. The Quakebot in Laird's approach is implemented in Soar, which is not BDI-based. Again, BDI agents like presented in this paper are more appropriate for the generation of folk psychological explanations than Soar agents.

Aylett and Louchart presented an account of intelligent agents with theory of mind which is based on simulation theory [1]. The agents are implemented in the emotionally driven agent architecture FAtiMA, in which agents assess the emotional impact of events in the world around them when deciding on their own actions. The agent's own mind is used to simulate what other agents might feel as a result of a possible action, and based on that the agent determines its own actions. The difference between this approach and the simulation-based approach presented in this paper is that the first focuses more on the emotional impact of behavior, and the latter focuses mostly on intentional behavior.

7. Conclusion

In the paper, two approaches for modeling agents with a theory of mind were introduced, based on the theory-theory and the simulation theory of mind. A case study was performed to compare agents with no theory of mind, a theory-theory of mind and a simulation theory of mind in an actual training scenario. It was found that all agent types were able to display the specified behavior, 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 is no difference between the two theory of mind approaches, but from a developer's perspective, the simulation theory has several advantages over the theory-theory approach. The simulation theoretic approach makes it easier introduce realistic errors in agent behavior due to an impaired theory of mind, and it promotes the reuse of code.

In future work, the use of agents with a theory of mind will be validated in a user study. In such an experiment, human subjects interact with agents in a virtual training, and after the training their performance on the training task is measured. The experiment knows two conditions, one condition in which the agents do have a theory of mind and another condition in which they do not. Subsequently, the performances of the subjects in both conditions can be compared to each other. The hypothesis is that subjects in the condition where the agents do have a theory of mind will perform better than subjects in the other group. When the results of such a study confirm the hypothesis, it is demonstrated that agents with a theory of mind can contribute to trainees' learning performances.

Another direction for future research is to explore the use of agents in virtual training with other social capacities besides a theory of mind. Agents could for instance be extended with emotions, such that they (partly) guide their decisions by emotions, and are able to express them, e.g. such as in [39]. Other examples are the ability to display cultural-aware behavior, norm-aware behavior [15], social responsible behavior [43]. In general, it is important to model those capacities in agents that are relevant to the training tasks. The more realistic the agents behave on these aspects, the more effective virtual training will become.

Acknowledgments

The authors would like to thank Michal Čáp and Mehdi Dastani, for their contribution to extending modular 2APL. Thanks to the anonymous reviewers for their constructive remarks. 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).

References

[1] R. Aylett and S. Louchart, If I were you: double appraisal in affective agents, *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*, IFAAMAS, 2008, pp. 1233-1236.

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

[3] G. Boella and L. Van der Torre, Groups as agents with mental attitudes, *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, IEEE Computer Society, 2004, pp. 964-971.

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

[5] T. Bosse, Z. Memon and J. Treur, A recursive BDI-agent model for theory of mind and its applications, *Applied Artificial Intelligence*, **25**(1) (2011), 1-44.

[6] M. Bratman, *Intention, plans and practical reason*, Harvard University Press, Cambridge, Massachusets, 1987.

[7] L. Braubach, A. Pokahr and W. Lamersdorf, Extending the capability concept for flexible BDI agent modularization, *Proceedings of ProMAS 2005*, Springer, 2005, pp. 139-155.

[8] P. Busetta, N. Howden, R. Rönnquist and A. Hodgson, Structuring BDI agents in functional clusters, *Proceedings of the 6th International Workshop on Intelligent Agents VI, Agent Theories, Architectures, and Languages*, Springer-Verlag, 2000, pp. 277-289.

[9] P. Busetta, R. Rönnquist, A. Hodgson and A. Lucas, Jack intelligent agents - components for intelligent agents in Java, *AgentLink News Letter*, 1999.

[10] P. Carruthers and P. Smith, Introduction, *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, pp. 1-10.

[11] P. Carruthers, Simulation and self-knowledge: a defence of the theory-theory, *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, pp. 22-38.

[12] B. Chrandrasekaran and J. Josephoson, Cognitive modeling for simulation goals: a research strategy for computer generated forces, *Proceedings of the 8th Computer Generated Forces and Behavioural Representation Conference*, 1999, pp. 239-250.

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

[14] M. Dastani, Modular rule-based programming in 2APL, *Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches*, A. Giurca, D. Gasevic and K. Taveter, eds., IGI Global, 2009, pp. 25-49.

[15] A. Doniec, R. Mandiau, S. Piechowiak and S. Espié, Controlling non-normative behaviors by anticipation for autonomous agents, *Web Intelligence and Agent Systems*, IOS Press, **6**(1) (2008), 29-42.

[16] R. Flin and K. Arbuthnot, eds., *Incident command: tales from the hot seat*, Ashgate Publising, 2002.

[17] P. Gmytrasiewicz and E. Durfee, A rigorous, operational formalization of recursive modeling, *Proceedings of the 1st International Conference on Mulitagent Systems*, AAAI Press, 1995, pp. 125-132.

[18] A. Goldman, In defence of the simulation theory, *Mind and Language*, **7**(1-2) (1992), 104-119.

[19] D. Gomboc, S. Solomon, M. G. Core, H. C. Lane and M. van Lent, Design recommendations to support automated explanation and tutoring, *Proceedings of the 14th Conference on Behav-*

ior Representation in Modeling and Simulation, Universal City, 2005.

[20] R. Gordon, 'Radical' simulationism, *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, pp. 11-21.

[21] M. Harbers, K. Van den Bosch and J.-J. Meyer, A methodology for developing self-explaining agents for virtual training, *Proceedings of Languages, Methodologies, and Development Tools for Multi-Agent Systems*, 2009, pp. 168-182.

[22] M. Harbers, K. Van den Bosch and J.-J. Meyer, Design and Evaluation of Explainable BDI Agents, *Proceedings of International Conference on Intelligent Agent Technology - Volume 2*, IGI Global, 2010, pp. 125-132.

[23] M. Harbers, K. Van den Bosch, and J.-J. Meyer, Agents with a theory of mind in virtual training, *Multi-Agent Systems for Education and Interactive Entertainment: Design, Use and Experience*, 2011, pp. 172-187.

[24] S. Harmon, D. Hoffmann, A. Gonzalez, R. Knauf and V. Barr, Validation of human behavior representation, *Proceedings of Workshop on Foundations for Verification and Validation (VV) in the 21st Century*, Society for Modeling and Simulation International, 2002, pp. 1-34.

[25] J. Heal, Simulation, theory, and content, *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, pp. 75-89.

[26] A. Heuvelink, *Cognitive models for training simulations*, Ph.D. Dissertation, VU University Amsterdam, 2009.

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

[28] B. Keysar, S. Lin and D. Barr, Limits on theory of mind use in adults, *Cognition*, **89**(1) (2003), 25-41.

[29] J. Laird, It knows what you're going to do: Adding anticipation to a Quakebot, *Proceedings of the 5th International Conference on Autonomous Agents*, 2001, pp. 385-392.

[30] W.L. Johnson, Agents that learn to explain themselves, *Proceedings of the 12th National Conference on Artificial Intelligence*, 1994, pp. 1257-1263.

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

[32] J. Perner, Simulation as explicitation of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix, *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, pp. 90-104.

[33] C. Peters, Foundations of an agent theory of mind model for conversation initiation in virtual environments, *Proceedings of AISB 2005 symposium on Virtual Social Agents*, 2005, pp. 163-170.

[34] A. Pokahr, L. Braubach and W. Lamersdorf, *Jadex: A BDI Reasoning Engine*, Kluwer Book, 2005.

[35] D. Premack and G. Woodruff, Does the chimpanzee have a theory of mind?, *Behavioral and Brain Sciences*, **1**(4) (1978), 515-526.

[36] D. Pynadath and S. Marsella, PsychSim: modeling theory of mind with decision-theoretic agents, *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005, pp. 1181-1186.

[37] A. Rao and M. Georgeff, Modeling rational agents within a BDI-architecture, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann publishers Inc., 1991, pp. 473-484.

[38] B. Scassellati, Theory of mind for a humanoid robot, *Autonomous Robots*, **12**(1) (2002), 13-24.

[39] B. Steunebrink, M. Dastani and J.-J. Meyer, Emotions to control agent deliberation, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, IFAAMAS, 2010, pp. 973-980.

[40] K. Van den Bosch, M. Harbers, A. Heuvelink and W. Van Doesburg, Intelligent agents for training on-board fire fighting, *Proceedings of the 2nd International Conference on Digital Human Modeling*, Springer, 2009, pp. 463-472.

[41] M. Van Lent, W. Fisher and M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, *Proceedings of IAAA 2004*, AAAI Press, 2004, pp. 900-907.

[42] H. Wimmer and J. Perner, Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception, *Cognition*, **13**(1) (1983), 103-128.

[43] M. Xu, L. Padgham, A. Mbala and J. Harland, Tracking Reliability and Helpfulness in Agent Interactions, *Web Intelligence and Agent Systems*, IOS Press, **5**(1) (2007), 31-46.

[44] J. Yen, X. Fan and R. Volz, Information Needs in Agent Teamwork, *Web Intelligence and Agent Systems*, IOS Press, **2**(3) (2004), 231-248.

[45] M. Young, Human performance model validation: one size does not fit all, *Proceedings of the 2003 Summer Computer Simulation Conference*, 2003, pp. 732-736.