M. Harbers, K. van den Bosch and J.-J. Meyer. (2011) Agents with a Theory of Mind in Virtual Training. In M. Beer, M. Fasli, and D. Richards (Eds.) Multi-Agent Systems for Education and Interactive Entertainment: Design, Use and Experience, pp. 172-187.

Chapter 9
Agents with a Theory of Mind in Virtual Training

Maaike Harbers
Utrecht University, The Netherlands

Karel van den Bosch
TNO Human Factors, The Netherlands

John-Jules Meyer
Utrecht University, The Netherlands

ABSTRACT

Virtual training provides an effective means to train complex, dynamic tasks like social interaction, negotiation and crisis management. The virtual characters with whom the trainee interacts are often played by autonomous, intelligent agents. For effective training, it is required that the agents behave in a believable way. In order to display believable social behavior, the agents must be able to take others' perspectives into account. This can be achieved by equipping them with a theory of mind, that is, the ability to attribute mental states such as beliefs and desires to others. In this chapter the authors describe an executable model for agents with a theory of mind, based on the BDI (belief desire intention) approach. The aim of the model is to develop agents that display believable social behavior and provide explanations about their behavior.

INTRODUCTION

Virtual training provides the opportunity to obtain realistic experiences in a safe environment, and is effectively employed for training complex, dynamic tasks like social interaction, negotiation or crisis management. To learn such tasks, trainees interact with virtual characters playing all kinds of roles, like classmate, negotiation partner, colleague, and team member. In virtual training, these roles can be played by humans, but also by intelligent agents. Trainees learn most from realistic human behavior as they are being prepared for interaction in the real world. Humans naturally display such behavior. However, people who can play the specialist roles are hard to find and not always available. Therefore, some trainees currently do not get enough training opportunities. When using intelligent agents instead of humans,

DOI: 10.4018/978-1-60960-080-8.ch009

less instructors are needed to train one trainee, and trainees will get more training opportunities.

Many researchers in artificial intelligence have tried to develop agents that seem to think, feel and live. If agents are to interact with other agents or humans, as in virtual training, they have to be able to display social behavior. Castelfranchi defined social behavior as taking others into account, and considering them as intentional beings (Castelfranchi, 1998). In other words, a social agent must be able to adopt the *intentional stance* towards others, i.e. to interpret others' behavior in terms of underlying mental concepts such as beliefs, goals and intentions (Dennett, 1987). The ability to attribute mental states to others and to understand that those can be different from one's own is called *theory of mind* (Whiten, 1991). To develop intelligent agents that display realistic social behavior, we will describe an executable model for agents with a theory of mind in this chapter.

Normally, neither humans nor agents have direct access to the mental states of others, so a theory of mind involves the *interpretation* of the observable behavior of others. Consequently, mental states attributed to another agent may be incorrect. Attributing incorrect knowledge and intentions to others is a well described phenomenon in cognitive sciences. Nickerson gives an extensive overview of research demonstrating the human tendency to overly impute one's own knowledge to others (Nickerson, 2003), and Keysar et al's research suggests human limits on the deployment of theory of mind in practical situations (Keysar et al, 2003). Agents with a theory of mind can improve trainees' performances by making them aware of their own and others' (limitations in) theory of mind use.

Extending intelligent agents with a theory of mind can contribute to virtual training in two ways: by their behavior and by explanations about their behavior. First, agents that act on the basis of a theory of mind give trainees the opportunity to experience how their behavior is interpreted by others. For instance, a trainee could respond slowly to an alarm call because a serious incident was reported and the trainee takes time to select a good strategy. However, the trainee's behavior might be interpreted by others as if there is no hurry, and in reaction, they will prepare themselves too slowly for handling the incident. Second, a trainee can learn more about how other agents interpreted his behavior by explanations that reveal the agents' attributed mental states. In our example, the agents could explain their behavior (preparing slowly) by their belief that the trainee believed that the incident was not serious. And when the agents explain their actions in terms of their mental states, the trainee can also check whether the beliefs and goals he attributed to the agents equal their actual ones. Explanations about the behavior of virtual characters yield better understanding of experienced training sessions, and make the trainee more alert to possible mistakes in the future.

Virtual training can be made more challenging by limiting an agent's theory of mind on purpose. With a limited theory of mind, the agent will attribute incorrect beliefs and goals to others and thereupon based the agent may select less appropriate actions. The trainee is challenged to detect the agent's incorrect attributions as early as possible by observing its behavior, and overcome possible problems. In that way, the trainee learns to cope with people that make false assumptions about others' beliefs and goals. On the other hand, agents can also employ their theory of mind to support the trainee. For instance, if they believe, based on their interpretation of the trainee's behavior, that the trainee misses certain information or is not aware that he has to perform a certain task, they can decide to help the trainee by warning him or taking over his tasks.

To summarize, in this chapter we will present an executable model for agents that are able to display believable social behavior and provide explanations about their behavior. We do that by extending existing agent models with a theory of