Guidelines for Developing Explainable Cognitive Models[1]

Maaike Harbers[1,3], Joost Broekens[2], Karel van den Bosch[3], John-Jules Meyer[1]

[1]Utrecht University, The Netherlands; [2]TU Delft, The Netherlands; [3]TNO Human Factors, The Netherlands
{maaike,jj}@cs.uu.nl, joost.broekens@gmail.com, karel.vandenbosch@tno.nl

Abstract

Cognitive models can be used to generate the behavior of virtual players in simulation-based training systems. To learn from such training, the virtual players must display realistic human behavior, and trainees need to understand why the other players behave the way they do. This understanding can be achieved by explaining the underlying reasons for the virtual players' behavior. In this paper, it is discussed how to design cognitive models in such a way that they are able to explain the behavior they generate. Three users studies were carried out to assess what type of explanations are useful for training, and how that relates to cognitive model design. Several guidelines for developing explainable cognitive models are proposed.

Keywords: Explanation, Cognitive modeling, Task analysis, Virtual training.

Introduction

Virtual training systems are increasingly used for training of complex tasks such as fire-fighting, crisis management, negotiation and social skills. To create valuable learning experiences, the virtual characters in the training scenario, e.g. the trainee's colleagues, opponents or team members, must display realistic behavior. Realistic behavior can be ensured by letting humans play these roles. However, the characters in virtual training systems often have specialist tasks which can only be played by experts, and human experts are often scarcely available. Alternatively, required human behavior can be represented in cognitive models, which gives trainees the opportunity to train whenever and wherever they like (Heuvelink, 2009).

A valuable learning experience requires more than interaction with virtual players displaying realistic behavior. To learn from training, trainees must (eventually) understand the behavior of the other players. Instructors can explain the motives behind other players' behavior, but that would reintroduce the availability problems with experts just mentioned. Preferably, cognitive models representing human behavior also have the ability to explain that behavior.

There are several systems providing explanations about non-human player behavior in virtual training systems, e.g. Debrief (Johnson, 1994), XAI I (Van Lent, Fisher, & Mancuso, 2004) and XAI II (Gomboc, Solomon, Core, Lane, & Lent, 2005; Core et al., 2006). However, none of these systems obtain their explanations directly from the cognitive models of virtual players. The XAI I system only provides explanations about the physical states of virtual players, e.g. their location and health. Debrief determines what must have been the beliefs of a virtual player, but does not have access to its actual beliefs. XAI II gives explanations in terms the underlying motivations of virtual players if those are represented in simulation, but this is often not the case. Moreover, as far as we know, the explanations of these systems are not empirically evaluated.

We advocate an approach that connects behavior generation and explanation. In other words, the cognitive models used to generate behavior can also be used to explain that behavior. The models are not necessarily similar to human reasoning, as long as they generate useful explanations. In this paper, we discuss three explorative studies in which users evaluate explanations generated by explainable cognitive models on their usefulness for learning. Based on the results, we present guidelines for designing explainable cognitive models.

The paper is organized as follows. First, we discuss what is known about how people explain behavior. Second, we introduce an approach for explainable cognitive models. Then, we describe three user studies evaluating explanations of these models, and discuss the results. From this discussion, we abstract guidelines for modeling and explaining virtual player behavior. We end the paper with a conclusion and suggestions for future research.

Explaining behavior

Keil provides an extensive overview of explanation in general, in which he categorizes explanations according to the causal patterns they employ, the *explanatory stances* they invoke, the domains of phenomena being explained, and whether they are value or emotion laden (Keil, 2006). Humans usually understand and explain their own and others' behavior by adopting the *intentional stance*.

Dennett distinguishes three explanatory stances: the mechanical, the design, and the intentional stance (Dennett, 1987). The mechanical stance considers simple physical objects and their interactions, the design stance considers entities as having purposes and functions, and the intentional stance considers entities as having beliefs, desires, and other mental contents that govern their behavior. The intentional stance is closely related to the notion of *folk psychology*. Folk psychology refers to the way people think that they think, and determines the language they use to describe their reasoning about actions in everyday conversation (Norling, 2004).

Attribution theory is one of the most important theories on people's behavior explanations, and focuses on the vari-

[1]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).

ous causes that people assign to events and behavior (Heider, 1958; Kelley, 1967). External attribution assigns causality to factors outside of the person, e.g. the weather. Internal attribution assigns causality to factors within the person, e.g. own level of competence. Related to attribution theory is the concept of *explanatory style*, i.e. people's tendency to explain causes of events in particular ways (Buchanan & Seligman, 1995). People with a negative explanatory style believe that positive events are caused by things outside their control and that negative events are caused by them. People with a positive explanatory style, in contrast, believe that positive events happened because of them and that negative events were not their fault. Explanatory style is part of someone's personality.

Attribution theory is criticized for not making a distinction between the explanation of intentional and unintentional behavior (Malle, 1999). In reaction, Malle provided a framework with different explanation modes. One explanation mode considers explanations about unintentional behavior, and three explanation modes consider explanations about intentional behavior: reason, causal history, and enabling factors explanations. Reason explanations are most often used and consist of beliefs and goals, causal history explanations explain the origin of beliefs and goals, and enabling factors explanations consider the capabilities of the actor.

A lot of research on explaining computer program behavior has been done in the field of expert systems (Swartout & Moore, 1993). Usually, outcomes like diagnoses or advices are explained by the steps that lead to it, e.g. the rules that were applied. It was found that the purpose of explanation has to be taken into account during system design. The information needed in explanations must be present, even though not necessary for the generation of behavior.

Putting these findings into the perspective of cognitive modeling and virtual training: trainees should get to understand the intentional behavior of virtual players. Different explanation theories use different terms for people's explanations of (intentional) human behavior. But whether called intentional, folk or reason explanations, they all refer to explanations in terms of mental concepts like beliefs, intentions and goals. Furthermore, when a cognitive model has to determine the behavior of a virtual player, it must be executable, e.g. by implementing the model in a cognitive architecture. From explanation research on expert systems we learned that the concepts needed for explanation must be present in the design. Consequently, to develop explainable cognitive models, concepts like motivations, beliefs, and goals need to be explicitly represented in the model.

An explainable cognitive model

Virtual players in training systems usually perform relatively well defined tasks. We therefore represent their behavior in the form of task hierarchies. Hierarchical task analysis is a well established technique in cognitive task analysis, and connects internal reasoning processes to external actions (Schraagen, Chipman, & Shalin, 2000). A task hierarchy has one main task, which is divided into subtasks, which are divided into subtasks, etc. Subtasks that are not divided are actions that can directly be executed in the environment. Adoption conditions are connected to each subtask, specifying the conditions under which a subtask can be adopted. Sardina et al pointed out the similarities between task hierarchies and BDI (Belief Desire Intention) models (Sardina, De Silva, & Padgham, 2006). The tasks and adoption conditions in a task hierarchy can be seen as goals and beliefs, respectively (see Figure 1). In earlier work we have elaborated the use of goal hierarchies for the representation virtual player behavior, and shown how these models can be implemented in a BDI (Beliefs Desire Intention) architecture, and thus be made executable (?, ?).

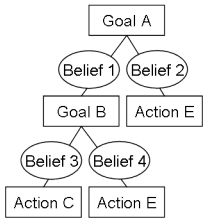

Figure 1: Example of a goal hierarchy.

There are four goal-subgoal relations: an *all* relation means that all subgoals must be achieved to achieve a goal, *one* means that exactly one subgoal must be achieved to achieve a goal, *seq* means that all subgoals must be achieved in a particular order to achieve a goal, and *if* means that a subgoal must only be achieved under certain conditions, i.e. when the player has certain beliefs. These relations yield different *action types*, i.e. the relation of an action to its parent goal.

An action can be explained by the goals and beliefs responsible for that action. However, providing the whole trace of beliefs and goals delivers long explanations with irrelevant information (Keil, 2006), in particular, with big goal hierarchies. Instead, a selection of 'explaining elements' can be provided to the trainee. For example, Action C in Figure 1 could be explained by Goal B, Goal A, belief 3, belief 1 or Action E (provided that E must follow C). More general, an action can be explained by different *explanation types*, respectively, the goal directly above an action (G+1), the goal two levels above an action (G+2), the beliefs one level above an action (B+1), the beliefs two levels above an action (B+2), and the goal or action that will be achieved after an action (Gnext).

Theories on human behavior explanation do not describe which explaining mental concepts should be part of an explanation. Malle's framework, for instance, does distinguish beliefs and goals in reason explanations, but does not (yet) describe in which situations which type is used more often (Malle, 1999). We performed three user studies to investigate which explanation types are considered useful to increase understanding of the training task. In particular, we investigated which explanation type is preferred for which action type. Our hypotheses are related to explanation stance,

length and type: 1) explanations in terms of beliefs and goals are appropriate for explaining virtual player behavior, 2) preferred explanations are relatively short and contain a selection among explaining beliefs and goals, and 3) preferred explanation type depends on the type of the action to be explained.

Three user studies

In this section we will give overviews of Study 1 (Harbers, Bosch, & Meyer, 2009b), 2 (Harbers, Bosch, & Meyer, 2010) and 3 (Broekens et al., 2010), and then discuss the results together. Only the results that are relevant for the discussion in this paper are presented. In all studies, the subjects were provided with a training scenario, and then asked to provide, select or judge explanations for several of the actions of the player(s) in the scenario. The independent variable in the studies is action type (actions with an *all*, *seq*, *one* or *if* relation to their parent) and the dependent variable is preferred explanation type (G+1, G+2, B+1, B+2, or Gnext). The explanations presented to the subjects were generated by implemented cognitive models of the virtual players.

Study 1: Onboard firefighting

Domain and task. The domain was onboard firefighting. The role to be trained was that of Officer of the Watch (OW), the person in command when there is a fire aboard a ship.

Subjects. The subjects (n=8) were instructors of the Royal Netherlands Navy and all expert on the training task.

Material. We used the CARIM system, a virtual training system for onboard firefighting (Bosch, Harbers, Heuvelink, & Van Doesburg, 2009). Three of the characters in the training scenario were modeled and implemented. The implementation was done in the programming language 2APL (Dastani, 2008). Questionnaires were administered to the subjects.

Procedure. Subjects played one scenario (approx 20 minutes), using the CARIM system, in which they were confronted with a fire aboard a Navy ship. Subsequently, they received a list with 12 actions of players in the scenario, and were asked to explain them in a way they considered useful for increasing trainees' understanding. Then, they received the same list of 12 actions, this time with four explanation alternatives (G+1, G+2, B+1, B+2) for each action. The subjects were asked to indicate which of the alternatives they considered most useful for increasing trainees' understanding.

Results. Regarding the first part of the questionnaire, we counted the number of elements in each of the subjects' own explanations, where an element is a goal, a fact, etc. Of the 88 explanations in total, 62 contained 1 element and 26 contained 2 elements. Furthermore, we categorized the elements in the subjects' explanations in different mental concepts. We were able to categorize all elements as either a belief or a goal: 52 beliefs and 62 goals. Table 1 shows the results of the second part of the questionnaire, the multiple choice questions. The agreement among the subjects for these results differed per action: for 5 actions at least 75% of the subjects preferred the same explanation, for 6 actions at least 50%, and for 1 action there was no explanation which at least 50% of the subjects preferred.

Action type	**Explanation type**			
	G+1	G+2	B+1	B+2
All (3 actions)	33%	50%	13%	4%
Seq (9 actions)	51%	21%	28%	0%

Table 1: Percentages of preferred explanation types per action type (n=8).

Study 2: Firefighting

Domain and task. The domain of this study was civil firefighting, and the role of the trainee was leading firefighter.

Subjects. The subjects (n=20) in Study 2 were unfamiliar to the training task. An advantage of non-expert subjects is that they do not have to imagine how useful the provided explanations are for understanding the training task. Instead, they can introspect to determine which explanations they consider useful. A disadvantage, on the other hand, is that non-experts cannot be expected to provide useful explanations for expert task actions themselves.

Material. A cognitive model of a leading firefighter was developed and implemented, again in 2APL. Questionnaires were used for the evaluation.

Procedure. The subjects were briefed about the training scenario, which involved a fire in a house. Subsequently, they received a list of 16 actions of the leading fire-fighter in the scenario with each four explanation alternatives (G+1, G+2, B+1, and Gnext). They were asked to indicate which explanation they considered most useful for understanding the task of leading fire-fighter.

Action type	**Explanation type**			
	G+1	G+2	B+1	Gnext
All (5 actions)	25%	16%	50%	9%
One (4 actions)	8%	8%	85%	0%
Seq (4 actions)	43%	14%	34%	10%
If (3 actions)	2%	2%	97%	0%

Table 2: Percentages of preferred explanation types per action type (n=20).

Results. Table 2 gives an overview of the results. For 7 of the actions at least 75% of the subjects preferred the same explanation, for 8 actions at least 50%, and for 1 action there was less than 50% agreement.

Study 3: Cooking

Domain and task. The domain of this study was cooking, and the training task was making pancakes. We purposely selected a simple training task, so that it was easy to find people

that could be considered experts.

Subjects. The subjects (n=30) were all familiar to this task.

Material. A cognitive model of a cook able to make pancakes was developed. The model was implemented in the programming language GOAL (Hindriks, 2009). Again, questionnaires were used for the evaluation.

Procedure. First, the subjects were briefed about the training scenario. Subsequently, they were asked to explain 11 of the cook's actions as they would to a student cook. Next, the subjects had to rate given explanations for all the 11 actions on their naturalness and usefulness on a scale of 1 to 5. The subjects were divided over condition 1, 2 and 3 in which they had to rate explanations of type G+1, B+1 and Gnext, respectively. In the last part of the questionnaire the subjects were shown the underlying goal hierarchy of the virtual player, and they were asked to indicate in the hierarchy by which beliefs and/or goals they would use to explain each of the 11 actions.

Results. The results of the subjects rating the usefulness of given explanations are shown in Table 3 (one of the actions was excluded from the analysis). The numbers are the average ratings of 10 subjects on 3 or 4 actions. The average

Action type	**Explanation type**		
	G+1	B+1	Gnext
All (3 actions)	3.2	2.5	3.4
One (3 actions)	3.0	2.4	2.0
Seq (4 actions)	2.9	2.8	1.8

Table 3: Average usefulness scores (scale 1-5) of action type per explanation type (n=30, n=10 per condition).

number of goals and/or beliefs that the subjects selected in the goal hierarchy for using in an explanation themselves was 1.7. One of the 30 subjects scored very high, and without this subject the average number of selected elements was 1.5.

Discussion

In this section we discuss the results of the user studies aiming to extract guidelines for developing and explaining cognitive models. The discussion is organized according to the three hypotheses concerning explanation stance, length and type.

From literature we learned that people adopt the intentional **explanatory stance** when they explain (intentional) human behavior. In other words, human(-like) behavior is explained by mental concepts such as beliefs and goals. The results of Study 1 show that it is possible to categorize the subjects' explanations in beliefs and goals, i.e. they are *compatible* with the intentional stance (we do not claim that this is the only way to categorize these explanations). In Study 3, the subjects' explanations were not categorized systematically, but an examination of the explanations provides a similar picture. Thus, the results confirm that people explain human-like virtual player behavior by the underlying beliefs and goals.

The results confirm our hypothesis that preferred explanations are relatively short. We expressed **explanation length** by the number of elements in an explanation, where an element is a fact, a goal, etc. In Study 1, the subjects' explanations had an average length of 1.3 elements, and in Study 3 the subjects selected an average of 1.7 elements from the goal hierarchy (1.5 if one outlier is eliminated from the data). The lower average in Study 1 might be due to the fact that the subjects had to write down complete explanations, whereas in Study 3 they only had to mark numbers of elements. So as expected, people's explanations about virtual player behavior usually only contain one or two elements.

As the results discussed so far confirm that explanations contain a selection of beliefs and goals, it makes sense to examine people's preferred **explanation type**. In Study 1, except for explanations of type B+2, all explanation types (G+1, G+2, B+1) were sometimes considered most useful by more than 50% of the subjects. In Study 2, for actions of type *one* and *if*, explanations containing a belief (B+1) were clearly preferred, and for actions of type *all* and *seq*, also explanations of other types (G+1 and G+2) were sometimes preferred by more than 50% of the subjects. These results are consistent with Study 1, in which only *all* and *seq* actions were examined. In Study 3, unlike Study 2, for all action types, explanations of type G+1 were on average rated higher than those of type B+1. Like in Study 2, for action types *one* and *seq*, Gnext explanations received relatively low ratings, and for actions of type *all*, they were highly rated. The usefulness of type Gnext explanations is closely related to the underlying cognitive model, which will be discussed in the next section. Interestingly, in the last part of Study 3, subjects often selected both a belief and a goal as their preferred explanation.

A remarkable difference between Study 1 and 3 on the one hand, and Study 2 on the other hand is that goal-based explanations were generally stronger preferred in the former, and belief-based explanations in the latter. A possible reason is that the subjects in Study 2 were unfamiliar, and those in Study 1 and 3 familiar with the training task. Data suggest that, on average, beliefs carry more idiosyncratic information and are harder to infer than goals (Malle, 1999). For subjects unfamiliar with a training task, belief-based explanations may provide more information underivable from the context than goal-based explanations. And expert subjects may not realize that goal-based explanations are easier to infer for trainees. Another explanation is that experts, more than non-experts, focus on the bigger picture of a virtual character's behavior. The subjects in Study 1 may be expected to know what would help trainees as they were instructors and had, besides being expert on the training task, didactical knowledge.

To conclude, action type is sometimes, but not always predictive for preferred explanation type. Of all studies, only Study 3 indicates to what extend explanations are preferred. The highest usefulness scores on action type *all*, *one* and *seq* are 3.4, 3.0 and 2.9, respectively. The scores are not low (all above the average of 2.5), but not very high either. In the experiments, we only provided subjects with explanations con-

taining one element, but the results seem to indicate that both beliefs and goals carry important information.

Modeling and explanation guidelines

Though the results of the three studies give no conclusive evidence, they provide directions for modeling and explaining virtual player behavior. In this section we present a set of guidelines for designing and explaining cognitive models.

The design and explanation of cognitive models are closely related in our approach. Though a virtual player's beliefs and goals remain unknown for users when a cognitive model is executed, they become visible when its behavior is explained. Thus, the elements in a cognitive model determine the content of its explanations. **Guideline**: the goals and beliefs in a goal hierarchy should be meaningful. Furthermore, two cognitive models with different underlying structures may display the same behavior, but generate different explanations. Figure 2, for instance, shows two possible positions of action E in a goal hierarchy. When both relations in this hierarchy are of the type *seq*, the position of action E does not effect the model's observable behavior, but it may influence they way it is explained, e.g. when explanations of the type G+1 are generated. Of course, developing a cognitive model always

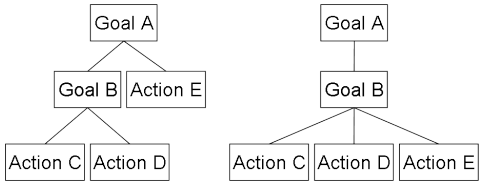

Figure 2: Same behavior, different explanations.

should be done with care, but as illustrated, this holds for explainable cognitive models in particular. **Guideline**: careful attention should be paid to the internal structure of the goal hierarchy. Though obvious, these two guidelines are crucial for developing useful explainable cognitive models.

In the previous section, we concluded that both beliefs and goals carry important information for explanations. The results showed that beliefs directly above an action (B+1) were considered most useful for explaining that action. Regarding goal-based explanations, the studies are less conclusive; several goal-based explanation types were considered useful (G+1, G+2 and Gnext) for different actions. But all together, goal-based explanations of type G+1 were most often preferred and highest rated. Moreover, people tend to use explanation types B+1 and G+1 together. **Guideline**: explanations should contain the belief(s) B+1 and the goal G+1.

The guidelines presented so far are general for all action types and supported by the results of all three studies. More specific guidelines that take action type into account can improve the default explanations. In the remainder of this section we will propose two additional, more specific guidelines.

In some cases an explanation of type Gnext can be added to the default explanation of G+1 and B+1. In contrast to G+1 and G+2 explanations, Gnext explanations do not contain goals from a particular level above the action. The level of the Gnext goal depends on the relations in the goal hierarchy. Here again, the usefulness of a Gnext explanation strongly depends on the underlying cognitive model. Consider, for instance, the two goal hierarchies in Figure 3. Goal B and C can be modeled as two neighboring goals or as goal and subgoal, e.g. when goal A, B and C represent *Report to head officer*, *Go to the head officer* and *Report new information*, respectively. In the first case, achieving goal B enables the achievement of goal C, and in the latter, goal C is achieved by achieving B. In Study 3, Gnext explanations were consid-

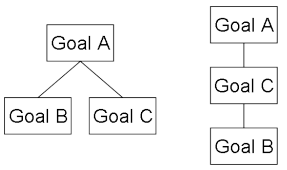

Figure 3: Neighbors or parent and sub-goal.

ered useful for actions of type *all*, where for all these *all* type actions it holds that their parents had a *seq* relation to their parents. **Guideline**: for actions of type *all*, when their parent goal has a *seq* relation, the explanation should contain the goal Gnext besides B+1 and G+1. Addition of a Gnext goal to the explanation may also be useful for other action types, but we have no evidence for that.

Another exception to the default rule concerns actions of the type *one*. The left side of Figure 4 represents a situation where action B is followed by action C or D, for example, the action *Take money* is followed by either *Cycle to the shop* or *Drive to the shop*. Action C and D are explained by goal A (G+1), e.g. *Buy ingredients*. However, a goal can only have one relation to its subgoal/actions, so the goal hierarchy in the left side is not allowed. The right side of Figure 4 shows how this situation should be represented. Goal A has a relation *seq* to its children, and a new goal X is introduced, e.g. *Go to the shop*, with a relation *one* to its children. Now, when action C and D are explained by their parent goal X, the explanation is not informative (I cycle to the shop because I want to go to the shop). In this case, it would be better to provide goal A as an explanation (I cycle to the shop because I want to buy ingredients). Although it may result in redundant goal-subgoal

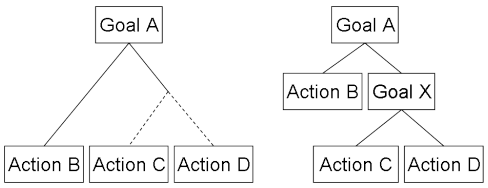

Figure 4: Explanation of actions with a one relation.

relations, we believe that from an explanation point of view a goal should have only one relation to its subgoals, as this simplifies interpretation of the cognitive model. **Guideline**: to explain actions of type *one*, instead of goal G+1, goal G+2 should be provided (i.e. B+1 and G+2).

Conclusion

In this paper we analyzed the results of three user studies investigating people's preferred explanations of virtual player behavior. From the analysis, we extracted a set of guidelines for developing and explaining cognitive models. In general, modeling should be done carefully, and by default, an action should be explained by the goal and belief directly above the action, i.e. explanation types G+1 and B+1. In addition, we introduced two guidelines for specific action types, which show how default explanations can be improved by providing extra or other elements in the goal hierarchy. More experimentation is needed for introducing more of these specific guidelines.

Another way to improve the explanations is by extending the cognitive model, for instance, by adding beliefs. Beliefs can contain information about the environment, e.g. resources that are available or events that just occured. Such beliefs are useful in particular and most often connected to *if* and *one* type actions. Beliefs can also contain information about internal reasoning processes, e.g. the given action is not yet executed, or a preceding action is executed. Such beliefs are more often connected to *all* and *seq* type actions. In these cases, it can be useful to add extra beliefs containing background information as adoption conditions. These background beliefs are always believed by the virtual player, so they do not effect the player's observable behavior, but they do add useful information to explanations.

There are many other directions in which this work can be extended. For instance, the cognitive models can be extended with emotions, a user model in which the trainee's knowledge is modeled can be used to select explanations, and the success of the approach in other domains can be examined. In future work we will first validate the present approach by comparing understanding of played training scenarios of trainees who did and did not receive explanations about virtual player behavior.

References

Bosch, K. Van den, Harbers, M., Heuvelink, A., & Van Doesburg, W. (2009). Intelligent agents for training on-board fire fighting. In *Proc. of the 2nd internat. conf. on digital human modeling* (p. 463-472). San Diago, CA: Springer Berlin/Heidelberg.

Broekens, J., Harbers, M., Hindriks, K., Bosch, K. Van den, Jonker, D., & Meyer, J.-J. (2010). *Do you get it? User evaluated explainable AI.* To appear.

Buchanan, G., & Seligman, M. (1995). *Explanatory style*. Erlbaum.

Core, M., Traum, T., Lane, H., Swartout, W., Gratch, J., & Van Lent, M. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation*, *82*(11), 685-701.

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

Dennett, D. (1987). *The intentional stance*. MIT Press.

Gomboc, D., Solomon, S., Core, M. G., Lane, H. C., & Lent, M. van. (2005). Design recommendations to support automated explanation and tutoring. In *Proc. of BRIMS 2005.* Universal City, CA..

Harbers, M., Bosch, K. Van den, & Meyer, J.-J. (2009a). A methodology for developing self-explaining agents for virtual training. In Decker, Sichman, Sierra, & Castelfranchi (Eds.), *Proc. of 8th int. conf. on autonomous agents and multiagent systems (aamas 2009)* (p. 1129-1130). Budapest, Hungary.

Harbers, M., Bosch, K. Van den, & Meyer, J.-J. (2009b). A study into preferred explanations of virtual agent behavior. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Vilhjlmsson (Eds.), *Proc. of IVA 2009* (p. 132-145). Amsterdam, Netherlands: Springer Berlin/Heidelberg.

Harbers, M., Bosch, K. Van den, & Meyer, J.-J. (2010). *Design and evaluation of explainable agents.* To appear.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley Sons.

Heuvelink, A. (2009). *Cognitive models for training simulations*. Unpublished doctoral dissertation, Vrije Universiteit Amsterdam, The Netherlands.

Hindriks, K. (2009). Multi-agent programming: Languages, tools and applications. In (p. 119-157). Springer.

Johnson, L. (1994). Agents that learn to explain themselves. In *Proc. of the 12th nat. conf. on artificial intelligence* (p. 1257-1263).

Keil, F. (2006). Explanation and understanding. *Annual Reviews Psychology*, *57*, 227-254.

Kelley, H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, p. 192-240). Lincoln: University of Nebraska Press.

Malle, B. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23-48.

Norling, E. (2004). Folk psychology for human modelling: Extending the BDI paradigm. In *Third internat. joint conf. on autonomous agents and multi agent systems* (p. 202-209). New York, USA.

Sardina, S., De Silva, L., & Padgham, L. (2006). Hierarchical planning in BDI agent programming languages: A formal approach. In *Proceedings of aamas 2006.* ACM Press.

Schraagen, J., Chipman, S., & Shalin, V. (Eds.). (2000). *Cognitive task analysis*. Mahway, New Jersey: Lawrence Erlbaum Associates.

Swartout, W., & Moore, J. (1993). Second-generation expert systems. In (p. 543-585). New York: Springer-Verlag.

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proc. of IAAA 2004.* Menlo Park, CA: AAAI Press.