A Study into Preferred Explanations of Virtual Agent Behavior

Maaike Harbers[1,2], Karel van den Bosch[2], and John-Jules Meyer[1]

[1] Utrecht University, P.O.Box 80.089, 3508 TB Utrecht, The Netherlands
{maaike,jj}@cs.uu.nl
[2] TNO Human Factors, P.O.Box 23, 3769 ZG Soesterberg, The Netherlands
karel.vandenbosch@tno.nl

Abstract. Virtual training systems provide an effective means to train people for complex, dynamic tasks such as crisis management or firefighting. Intelligent agents are often used to play the characters with whom a trainee interacts. To increase the trainee's understanding of played scenarios, several accounts of agents that can explain the reasons for their actions have been proposed. This paper describes an empirical study of what instructors consider useful agent explanations for trainees. It was found that different explanations types were preferred for different actions, e.g. conditions enabling action execution, goals underlying an action, or goals that become achievable after action execution. When an action has important consequences for other agents, instructors suggest that the others' perspectives should be part of the explanation.

1 Introduction

This paper presents a study about explanations of intelligent virtual agents about their own behavior. Several accounts for such self-explaining agents for virtual training have been proposed [7, 11, 4, 1]. In general, self-explaining agents act in virtual training systems used to train people for complex, dynamic tasks in which fast decision making is required, e.g. the persons in command in crisis management, military missions or fire-fighting. During a training session, a trainee interacts with the virtual agents, which play the role of e.g. team-member or opponent. After the training session is over, the agents can be queried or give explanations on their own initiative about their actions in the played session, aiming to give trainees better insight in the played training session.

Explanations exist in a wide variety of forms. Explanations of why one should wear seat belts, why trees grow, and why Anna was angry with John last week, each have different properties and are created according to different mechanisms. Even single phenomena or processes can be explained in different ways. For instance, the vase fell because of the gravity force, because Chris pushed it, or because Chris was distracted by his cat. As there are so many possible ways to explain phenomena, events and processes, explaining all facets is usually neither possible, nor desired [6]. Thus, in order to provide useful explanations one should

choose an explanation type and select the information that fits the domain and the people to whom the explanation is directed.

A way to categorize explanations is according to the *explanatory stance* to be adopted for framing the explanation [6]. Dennett distinguishes three explanatory stances: the mechanical, the design and the intentional stance [3]. 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. Humans usually understand and explain their own and others' behavior by adopting the intentional stance. Most accounts of self-explaining agents give explanations in terms of an agent's beliefs [7] or motivations [4, 1] that were responsible for its actions. We believe that the intentional stance distinguishes explanations of agents from explanations provided by expert systems, in which no intentionality is involved [12].

In earlier work we have proposed an account of self-explaining agents which is able to provide explanations in terms of beliefs and goals [5]. Though the scope of these agents' possible explanations is restricted by adopting the intentional stance, they can still explain one action in several ways. There are usually several mental concepts that underly one action, but not all of them are equally relevant in an explanation. Especially when the agent models are complex, providing all beliefs and goals underlying an action does not result in useful explanations. Instead, explanations containing a selection of the explaining mental concepts are probably more effective.

The purpose of the study presented in this paper is twofold. First, it serves to examine whether the explanatory stance we used in our approach for self-explaining agents is considered useful by instructors. We aim to find empirical indications that explanations which are considered useful by instructors are compatible with the intentional stance. We consult instructors' on what they consider useful explanations for trainees, as instructors have knowledge about both the task domain and didactic aspects. Second, we want to use the results of this study to further develop our approach of self-explaining agents, so that within the scope of possible explanations, useful ones are selected. We will consider several properties of explanations: explanation length, abstraction level and explanation type. The experiments in this paper aim to shed light on instructors' preferences on these aspects.

2 Methods

The subjects participating in the experiments had to play a training session involving several virtual agents. After the scenario was completed, the subjects were provided by possible explanations for actions performed by the virtual agents, and asked to select the explanation which they considered most useful for a trainee. In this section we will discuss the virtual training system that was used, the generation of possible explanations, and more details on the experimental setup.

2.1 Training on-board fire fighting

The subjects played a training session with the Carim system[1], a virtual training system developed for the Royal Netherlands Navy to train the tasks of an Officer of the Watch (for a more extensive overview of the system see [10]). The Officer of the Watch is the person who is in command when there is a fire aboard a navy frigate. From the Technical Center of the ship he collects information, makes an assessment of the situation, develops plans to solve the incident, instructs other people, monitors the situation, and adjusts his plans if necessary. The Officer of the Watch communicates with several other officers, of which the Chief of the Watch, the Leader Confinement Team, and the Leader Attack Team are the most important. In a typical incident scenario, the Officer of the Watch and the Chief of the Watch remain in the Technical Center, the Leader Attack Team is situated close to the location of the incident, and the Leader Confinement Team moves between both locations. One training session takes about half an hour to complete.

Fig. 1. A snapshot of the Carim system: communication with a virtual agent.

The Carim system is a stand-alone, low-cost desktop simulation trainer, to be used by a single trainee who is playing the role of Officer of the Watch. The trainee can freely navigate through the Technical Center. All equipment that the Officer of the Watch normally uses is simulated and available to the trainee, e.g. a map of the ship, information panels and communication equipment. Communication from agent to trainee happens by playing pre-recorded speech expressions, and a trainee can communicate with an agent by selecting speech acts from a menu (figure 1). These menus are agent-specific and may change over the course of a training session.

[1] The Carim system has been developed by TNO and VSTEP

The course of a training session in the Carim system is guided by a scenario script. The script defines for each possible situation what should happen, which is either an event in the environment or an action of an agent. The trainee has certain freedom to act the way he wants to act, but if he deviates from the storyline in the scenario, the simulation redirects the trainee back to the intended scenario. For instance, if it is necessary that the trainee contacts the Leader Attack Team, the Chief of the Watch will repeat an advice to contact the Leader Attack Team till the trainee does so. Currently, a new version of the Carim system is being developed in which the behavior of agents is not scripted, but generated online by intelligent agents. Advantages of intelligent agents are that they are able to deal with unexpected trainee behavior, and thus yield more freedom for the trainee and more diverse courses of a training scenario. Moreover, intelligent agents can more easily be reused in different scenarios. However, the improved version of the Carim system was not available yet at the time the experiments in this paper were performed.

2.2 Explanation generation by simulation

In the ideal case, the behavior of the virtual characters in the training system would be generated autonomously and online by self-explaining agents. Then, the agents would create logs about their decisions and actions during the scenario, and based on these logs, give explanations for their behavior in the scenario afterwards. However, because no connection between intelligent agents and virtual environment had been established yet, we had to obtain explanations of agents in another way. We did so by running a separate simulation with only agents, and no visualization of the environment. These agents were not scripted, but instead, generated behavior in an intelligent way. During the simulation, the self-explaining agents built up a log about their decisions and actions, and based on these logs we could derive explanations. We run the simulation before the actual experiment took place, and in the experiment we presented the beforehand obtained explanations to the subjects.

We have modeled and implemented three of the agents in the Carim scenario: the Chief of the Watch, the Leader Confinement Team, and the Leader Attack Team. While modeling the agents, we ensured that they would generate the same behavior as the scripted virtual agents with whom the subjects would interact. The behavior of the scripted agents was almost completely known because the scenario script of the Carim system only allows for little deviation from the intended storyline. The difference between modeled agents and the scripted agents is that the modeled agents make reasoning steps in order to generate behavior and the scripted agents do not. These reasoning steps are stored in a log, and from this log explanations can be derived. Because the actions of the scripted and modeled agents are equal, the derived explanations are exactly the same as when there would be a connection between agents and virtual environment.

We used the approach for developing self-explaining agents that we recently proposed [5]. In this approach, the conceptual model of an agent is a task hierarchy containing all its possible tasks. A task hierarchy representation language is

used to represent an agent's tasks, subtasks and the conditions for adopting and achieving tasks. It is specified how such a task hierarchy can be translated to an agent implemented in a BDI-based (Belief Desire Intention) agent programming language. The translation is based on similarities between task hierarchies and BDI models[9]. Tasks are implemented as goals, actions (tasks at the bottom of a hierarchy) are implemented as plans, and adoption conditions are implemented as beliefs.

Following this approach, we constructed task hierarchies of the Chief of the Watch, the Leader Confinement Team, and the Leader Attack Team in the Carim scenario, and implemented them in the agent programming language 2APL [2]. For the construction of the task hierarchies, we used task descriptions provided by the Navy and interviews with experts. Figure 2 shows a part of the task hierarchy of the Leader Attack Team agent. The task *Initiate fire attack*, for instance, has three subtasks: *Go to location*, *Develop a plan*, and *Instruct team*. Only for two tasks (*Initiate fire attack* and *Develop a plan*) the conditions under which they are adopted are shown, but all other tasks have adoption conditions as well. For instance, for achieving the task *Initiate fire attack* one can only adopt the task to *Develop a plan* when one is *At location*.

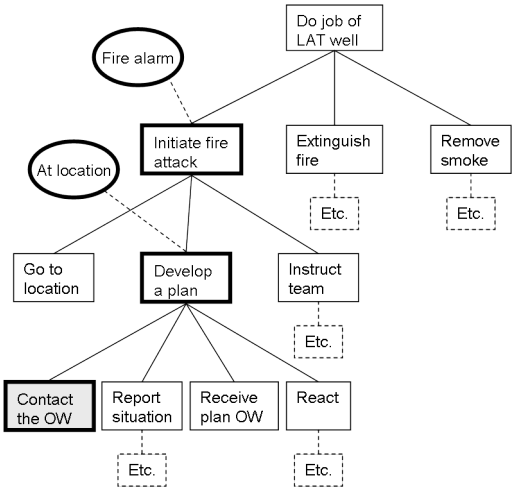

Fig. 2. Part of task hierarchy of the Leader Attack Team containing tasks (boxes) and adoption conditions (circles).

By developing agents according to this approach, the elements in the deliberation process generating behavior can also be used to explain that behavior, i.e. the goals and beliefs underlying an action also explain that action. In figure 2, for instance, the tasks and adoption conditions underlying and thus explaining the action *Contact the OW* are marked by a bold line. To enable agents to provide explanations, they need to have knowledge about their own internal structure,

which is realized by adding a representation of the agent's own task hierarchy to its belief base. Moreover, when the agent is executed, a log of the agent's actions is created in its belief base. With knowledge about its task structure and past actions, an agent can explain its actions by providing the beliefs and goals that were involved in the generation of that action.

We equipped the three Carim agents with explanations capabilities and run them. Figure 3 shows a part of the Leader Attack Team agent's belief base after it was run. The left part of the code represents the agent's beliefs about its task hierarchy, and the right part shows the log that was created during the execution.

```
task(DoJobWell,                                    log(t(1), goToLocation).
   [(InitiateFireAttack, alarm),                   log(t(2), contactOW).
   (ExtinguishFire, attackInitiated),              log(t(3), reportFire).
   (RemoveSmoke, fireExtinguished and smoke)]).    log(t(4), reportVictims).
task(InitiateFireAttack,                           etc.
   [(GoToLocation, not atLocation),
   (DevelopPlan, atLocation),
   (InstructTeam, planDevleoped)]).
etc.
```

Fig. 3. Part of the belief base of the Leader Attack Team after execution.

The left part of the code shows a representation of a part of the Leader Attack Team's task hierarchy. Each task in the hierarchy is represented by a belief containing information about the identity of the task, the identity of its subtasks, and the adoption conditions of the subtasks. For example, the task *DoJobWell* has three subtasks, *InitiateFireAttack*, *ExtinguishFire* and *RemoveSmoke*, with adoption conditions *alarm*, *attackInitiated*, and *fireExtinguished and smoke*, respectively. The right part of the code shows log beliefs containing information about the actions that were executed and the time of their execution. From these two types of information in the agent's belief base, several explanations for one action can be derived. For example, the Leader Attack Team went to the location of the incident because a) it had to do its job well, b) the alarm bell rang, c) it had to initiate the fire attack, or d) it was not at the location of the incident yet.

2.3 Experimental setup

We conducted two experimentation sessions in which subjects had to play a scenario in the Carim system and indicate their preferences about explanations of the behavior of the agents in the scenario. The subjects, 15 in total, were all instructors of the Dutch Navy with knowledge about the task domain, and experience in teaching. Before they started, the subjects received instructions about the objective of the experiments and what they were expected to do. None

of them had used the Carim system before. To get acquainted to the system, they received some time to practice with navigating their avatar, communicating with other agents, and marking an incident on a damage panel. The subjects were asked to fill in a questionnaire after playing the session.

The questionnaires consisted of two parts. In part I, the subjects were asked to explain 12 actions which had been performed by the virtual agents in the scenario, four actions of each of the three agents we modeled. For example, the subjects had to explain why the Leader Attack Team contacted the Officer of the Watch. They were instructed to provide explanations from the perspective of the agent that executed the action, and to keep in mind that the explanations should help the trainee to achieve a proper understanding of the situation. In part II, the same 12 actions were presented to the subjects, and this time they were asked to select the best out of four possible explanations. The explanations were derived from the agents' belief bases after the offline simulation with intelligent agents. We translated the programming code explanations to sentences in natural language by hand. For example, the subjects had to indicate their preferred explanation for the following question.

The Leader Attack Team contacted the Officer of the Watch because...
- *there was a fire alarm*
- *he wants to initiate the fire attack*
- *he arrived at the location of the incident*
- *he wants to develop an attack plan*

The four explanations are, respectively, an abstract belief, an abstract goal, a detailed belief and a detailed goal. Abstract explanations contain goals and beliefs higher in the agent's task hierarchy, and detailed explanations contain goals and beliefs higher and lower in the agent's task hierarchy. In part II, the subjects were again instructed to keep in mind that the explanations should increase trainees' understanding of the situation. The subjects had to give their own explanations before they saw the agents' explanations to ensure that their own explanations were not influenced by the explanations provided to them in the second part of the questionnaire.

The first session was conducted with 8 subjects, and mainly focused on the question whether explanations preferred by instructors are compatible with the intentional stance. Namely, the answers in part II of the questionnaire, in which subjects select one out of several intentional explanations, are only valuable if the instructors consider intentional explanations useful at all. The compatibility of the instructors' preferred explanations with the intentional stance was obtained by analyzing the instructors' own explanations in part I of the questionnaire. We expected that the instructors' preferred explanations would be compatible with the intentional stance.

The second experimentation session aimed to obtain more detailed information on the nature of preferred explanations, such as preferred explanation length, type and abstraction level. Concerning explanation length, we expected

that short explanations would be preferred over long ones because not all information that explains an action is relevant. Concerning explanation type, we expected that explanations containing a belief with the precondition for an action, or a goal to be achieved by an action would be preferred over other types of explanations. Finally, concerning preferred abstraction level, we expected that explanations containing detailed, low-level information would be preferred over explanations with abstract, high-level information.

3 Results

All of the subjects were able to solve the incident presented to them in the training scenario. Though some of the subjects had some difficulties with navigating their avatar, they generally rated the training system positively. Section 3.1 discusses the results of experimentation session 1, and presents data obtained from part I of the questionnaire. Section 3.2 discusses the results of experimentation session 2, and presents data obtained from part II of the questionnaire. Note that some of the data obtained in session 1 from part II of the questionnaire are also presented in section 3.2.

3.1 Session 1

The first experimentation session was conducted with 8 subjects. From part I of the questionnaire we obtained 88 explanations of virtual agent actions, provided by the subjects themselves. Note that in 8 occasions, a subject was unable to provide an explanation, as 8 times 12 should deliver 96 explanations.

Subjects' own explanations: explanation length. The first categorization of explanations is according to their length. We defined explanation length by the number of elements, where an element is a goal, a belief, a fact, etc. Table 1 shows the frequencies of the number of elements in the subjects' explanations. The results show that most explanations contained only 1 element (70%). All

Length	# explanations
1 element	62
2 elements	26
>2 elements	0

Table 1. Frequencies of the number of elements in the provided explanations (n=8).

others contained 2 elements (30%). No explanations with more than 2 elements were given.

Subjects' own explanations: explanation type. A second way to categorize the subjects' explanations is according to type. More specifically, explanation elements can be categorized according to type. Our aim was to examine whether the subjects' explanations are compatible with the intentional stance. We thus tried to map the provided explanation elements to intentional concepts such as beliefs, goals and intentions.

An examination of the provided explanations resulted into five types of explanation elements: the condition for executing an action, background information concerning an action, the goal to be achieved an the action, the goal that becomes achievable after executing an action, and others' goals that become achievable after executing an action. A *condition* for an action was for example 'I went to the location of the incident because I heard the alarm message'. An example of *background information* is 'the Officer of the Watch and the Leader Attack Team communicate by a headphone'. An explanation with a *goal to be achieved* is for instance 'I put water on the fire to extinguish it'. An explanation containing an *enabled goal* is e.g. 'I prepared fire hoses to extinguish the fire'. Finally, an example of an explanation in terms of an *other's goal* is 'if I make the location voltage free, my colleague can safely use water in the room'.

The first two types, condition and background information can be considered as beliefs, and the last three types, own goal, enabled goal and other's goal, are all goals. We do not claim that our classification is the only one possible. These results should rather be seen as an explorative examination of whether the provided explanations are compatible with the intentional stance. Table 2 shows the number of provided elements per explanation type. If an explanation contained two elements, e.g. a goal and background information, both elements were counted as a half. A remark about table 2 is that some of the explanations

Type	**# elements**
Belief (condition)	10
Background information	10
Goal	12.5
Enabled goal	34
Other's goal	21.5

Table 2. Number of explanations per explanation type (n=8).

classified as enabled goals could also be classified as goals. For instance, the explanation 'the Leader confinement team goes to the TC to report to the Officer of the Watch' can be classified in two ways. Namely, the explaining element 'to report to the Officer of the Watch' can be seen as a goal of which going to the TC is a subgoal, but also as an enabled goal that can be achieved after the Leader Confinement Team arrived in the TC. In the first interpretation the explanation would be classified as a goal, and in the second as an enabled goal. In case of such ambiguity, we have chosen for the second interpretation, and classified the explanation as an enabled goal.

In the first experimentation round, the second part of the questionnaire only contained explanations in terms of beliefs forming a condition and goals to be achieved by the action. However, the results in table 2 show that many of the explanations were in terms of enabled goals and others' goals. Therefore, we decided to add more possible explanations to the second part of the questionnaire. In figure 2 one can see that an explanation in terms of an enabled goal for the action 'The Leader Attack Team contacted the Officer of the Watch' is that 'The Leader Attack Team wants to report the situation to the Officer of the Watch'. Explanations in terms of others' goals cannot be derived from an agent's own belief base, but it is possible to look at task hierarchies of other agents and formulate explanations in terms of others' goals.

3.2 Session 2

The second experimentation session was conducted with 7 subjects. Part II of the questionnaire was adjusted by adding an explanation in terms of an enabled goal to the answers where possible, and adding an explanation in terms of an other's goal to the answers for all actions. As explanations in terms of others' goals were not derivable from the agents' own belief bases, preferences on this type of explanations were asked in a separate question.

Multiple choice: explanation type (3 choices) and abstraction level. There were five actions for which an explanation in terms of an enabled goal could be derived from the agents' task hierarchies, and for these actions the subjects could select one of five possible explanations in part II of the questionnaire. Table 3 shows for these actions which explanation type was preferred, and whether at least 75% or 50% of the subjects agreed on that. Thus, the italic numbers in the table are action numbers and not frequencies. The general agreement among

Type	**Abstraction**	**>75%**	**>50%**
Belief	Detailed	*8*	-
	Abstract	-	-
Goal	Detailed	-	*1*
	Abstract	-	*7*
Enabled Goal	-	*2,5*	-

Table 3. Preferred explanation types and abstraction levels for actions 1,2,5,7,8 (n=7).

the subjects expressed in a multi-rater kappa coefficient [8] was 0.55. The results show that for some actions (action 2 and 5) a large majority of the subjects preferred an explanation in terms of an enabled goal. However, explanations in terms of enabled goals were not always preferred. Subjects even agreed for more than 75% that action 8 could best be explained in terms of a detailed condition belief.

Multiple choice: explanation type (2 choices) and abstraction level. For the actions for which no explanation in terms of an enabled goal could be derived from the agents' belief bases, subjects could choose between four options. As these questions were equal to those in the first experimentation session, table 4 shows the results based on the answers of all 15 subjects. Out of seven actions,

Type	**Abstraction**	**>75%**	**>50%**
Belief	Detailed	*10*	*3*
	Abstract	-	-
Goal	Detailed	*9*	*11,12*
	Abstract	-	*6*

Table 4. Preferred explanation types and abstraction levels for actions 3,4,6,9,10,11,12 (n=15).

only for two actions (9 and 10) more than 75% of the subjects agreed on the preferred explanation type, which is reflected in a rather low kappa coefficient of 0.33. For action 4, no preference on which at least 50% of the sujects agreed was found.

Multiple choice: explanations involving other agents' perspectives. In the second experimentation session, the 12 multiple choice questions in part II were each followed up by a second question concerning explanations in terms of others' goals. After indicating their preference of four or five possible explanations, subjects were asked to compare their first answer to another (fifth or sixth) option. This extra option was an explanation in terms of an other's goal, for instance as follows.

The Leader Attack Team contacted the Officer of the Watch because...
- *< answer given in part a >*
- *the Officer of the Watch knows he is there*

The results of the follow up question are presented in table 5. A kappa of 0.43

Type	**>75%**	**>50%**
First choice	*1*	*7,8*
Other's goal	*4,5,9,10,11,12*	*2,3,6*

Table 5. Amount of explanations and the actions per explanation type (n=7).

for the overall agreement among the subjects was found. The results show that for 9 out of 12 actions, the subjects preferred explanations in terms of an other's goal over their first choice. For six of the actions (4,5,9,10,11,12) more than 75% of the subjects preferred an explanation in terms of an other's goal and only for one action (1) more than 75% of the subjects agreed on a preference for an

explanation not based on an other's goal. As explanations in terms of others' goals were not based on the belief bases of the agents we modeled, the data in table 5 should not be considered as final results, but as an exploration of possibly preferred explanations.

4 Discussion

The first objective of the study was to examine whether preferred explanations are compatible with an intentional perspective. In part I of the questionnaire, the subjects were asked to provide explanations without any constraints concerning explanation types to be used. We were able to classify the elements in the subjects' own explanations in five explanation types, which were all either belief-based or goal-based (table 2). Though the explanations might be classifiable in other ways, possibly in non-intentional explanation types, it was possible to understand the explanations from an intentional perspective. We may thus conclude that the preferred explanations are compatible with the intentional stance, and that using a BDI-based approach is an appropriate method for developing self-explaining agents.

In section 2.3, we formulated three expectations about the nature of preferred explanations. Concerning explanation length, we expected that short explanations would be preferred over long ones. In part I of the questionnaire, the subjects were asked to provide their own explanations, whereby no restrictions on explanation length were given. Table 1 showed that their explanations in most cases only contained one element, and never more than two elements. These results thus confirm our expectations. Consequently, self-explaining agents should make a selection of elements which they provide in an explanation.

Concerning preferred explanation type, we expected that most of the instructors' explanations about an action would either involve a belief with the condition enabling execution of the action, or a goal for which the action is executed. Though part of the instructors' explanations in part I could be classified into one of these categories, the results in table 2 show that three other explanation types were also used, namely background information, enabled goals and others' goals. Results from part II of the questionnaire confirmed the results of part I. Namely, table 3 and 5 show that the instructors sometimes selected other explanation types than the ones we originally expected. We may conclude that our expectations were partly supported by the results. Within our approach of self-explaining agents it was already possible to provide explanations in terms of enabled goals, but the agents should be extended with the capability to provide explanations in terms of others' goals as well.

Our last expectation involved the abstraction level of preferred explanations. We expected that detailed explanations would be preferred over abstract ones. In this study, detailed and abstract explanations consisted of mental concepts low (just above action level) and high in an agent's task hierarchy, respectively. We cannot give a general conclusion concerning preferred abstraction level because the data only give information about the preferred abstraction level of two types

of explanations, condition beliefs and own goals. For belief-based explanations, the results clearly show that detailed explanations are preferred over abstract ones (table 3 and 4). For goal-based explanations, the results in table 3 and 4 also show that detailed explanations are preferred over abstract ones, but not as convincingly as for belief-based explanations.

A possible explanation for the low score on abstract belief-based explanations is that condition beliefs are often directly related to events in the environment. Abstract beliefs take place earlier in time than detailed beliefs, for example, the fire alarm rings before the Leader Attack Team reaches the location of the incident, and it is plausible that more recent cues are preferred over older ones.

5 Conclusion and future work

In this paper we described a study of preferred explanations of virtual agents in a training context. Our goal with this study was twofold. First, it aimed to explore whether our BDI-based approach of self-explaining agents as described in section 2.2 is appropriate for the generation of explanations. The results of the experiments supported our expectation that preferred explanations of virtual agent behavior are compatible with an intentional perspective. Thus, explanations in terms of beliefs and goals are expected to enhance trainees' understanding of the training situations.

Second, the study was meant to obtain information on how our approach for developing self-explaining agents could be improved. An important finding was that for some actions, instructors preferred explanations in terms of enabled goals and others' goals. Explanations in terms of enabled goals can already be derived from the self-explaining agents' belief bases in the current approach. However, to generate explanations in terms of others' goals, the agents' models need to be extended with beliefs about other agents' goals. We are currently working on extending the self-explaining agents with a theory of mind, i.e. the agents are equipped with a 'theory' about the beliefs and goals of other agents. With this extension it should be possible to generate explanations that are based on other agents' goals.

Another outcome of the study was that different actions were explained in different ways. In future work we will examine different situations in which actions are executed in relation to their preferred explanation. For instance, only if an action has important consequences for other agents, an explanation in terms of others' goals may be preferred. Finding such relations will help to develop a mechanism that selects a useful explanation among possible explaining mental concepts.

After improving the approach of self-explaining agents, we plan to perform a new set of experiments. These experiments will not be meant to explore, as the study described in this paper, but rather to validate the approach. Consequently, the subjects of the experiments will be trainees instead of instructors. The results of such a study can give insight on whether the explanations really enhance trainees' understanding of the training situations.

Acknowledgments. 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. M. Core, T. Traum, H. Lane, W. Swartout, J. Gratch, and M. van Lent. Teaching negotiation skills through practice and reflection with virtual humans. *Simulation*, 82(11):685–701, 2006.
2. M. Dastani. 2APL: a practical agent programming language. *Autonomous Agents and Multi-agent Systems*, 16(3):214–248, 2008.
3. D. Dennett. *The Intentional Stance*. MIT Press, 1987.
4. D. Gomboc, S. Solomon, M. G. Core, H. C. Lane, and M. van Lent. Design recommendations to support automated explanation and tutoring. In *Proc. of the 14th Conf. on Behavior Representation in Modeling and Simulation*, Universal City, CA., 2005.
5. M. Harbers, K. Van den Bosch, and J. Meyer. A methodology for developing self-explaining agents for virtual training. In Decker, Sichman, Sierra, and Castelfranchi, editors, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 1129–1130, Budapest, Hungary, 2009.
6. F. Keil. Explanation and understanding. *Annual Reviews Psychology*, 57:227–254, 2006.
7. W. Lewis Johnson. Agents that learn to explain themselves. In *Proc. of the 12th Nat. Conf. on Artificial Intelligence*, pages 1257–1263, 1994.
8. J. Randolph. Online kappa calculator. Retrieved March 6, 2009 , from http://justus.randolph.name/kappa, 2008.
9. S. Sardina, L. De Silva, and L. Padgham. Hierarchical planning in bdi agent programming languages: A formal approach. In *Proceedings of AAMAS 2006*. ACM Press, 2006.
10. K. Van den Bosch, M. Harbers, A. Heuvelink, and W. Van Doesburg. Intelligent agents for training on-board fire fighting. To appear, 2009.
11. M. Van Lent, W. Fisher, and M. Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proc. of IAAA 2004*, Menlo Park, CA, 2004. AAAI Press.
12. R. Ye and P. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly*, 19(2):157–172, 1995.