Explanation in Human-Agent Teamwork

Maaike Harbers[1], Jeffrey M. Bradshaw[2], Matthew Johnson[2], Paul Feltovich[2], Karel van den Bosch[3] and John-Jules Meyer[4]

[1] TU Delft, P.O. Box 5031, 2600 GA, Delft, The Netherlands
m.harbers@tudelft.nl
[2] IHMC, 40 South Alcaniz, Pensacola, FL 32502, United States
{jbradshaw,mjohnson,pfeltovich}@ihmc.us
[3] TNO, P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands
karel.vandenbosch@tno.nl
[4] Utrecht University, P.O. Box 80.089, 3508 TB, The Netherlands
j.j.c.meyer@uu.nl

Abstract. There are several applications in which humans and agents jointly perform a task. If the task involves interdependence among the team members, coordination is required to achieve good team performance. This paper discusses the role of explanation in coordination in human-agent teams. Explanations about agent behavior for humans can improve coordination in human-agent teams for two reasons. First, with more knowledge about an agent's actions and plans, humans can more easily adapt their own behavior to that of the agent. Second, with more insight in the reasons behind an agent's behavior, humans will have more trust in the agents, and thefefore more easily coordinate their actions. The paper also presents a study in the BW4T testbed that examines the effects of agents explaining their behavior on human-agent team performance. The results of this study show that explanations about agent behavior do not always lead to better team performance, but they do impact the user experience in a positive way.

1 Introduction

When the members of a team jointly perform a task, it often happens that one team member is dependent on other team members for achieving a subtask. For instance, it may happen that a team member can only start to achieve subtask A after someone else has achieved subtask B, or that a team member can only start to achieve subtask C after another team member has started to achieve subtask D. When team members are dependent on each other for achieving a task, they are *interdependent* [16]. Interdependent team members need to coordinate their actions in order to achieve a good team performance. Namely, it is ineffecient when two team members are separately trying to achieve a task that can easily be achieved by one team member. Thus, the better the actions of different team members are coordinated, the higher the performance of the team will be.

Coordination of actions is not only important in human teams, but also in teams that consist of a mix of humans and software agents. Therefore, when

developing agents that aim to participate in a human-agent team, it is important to make them able to coordinate their actions with other team members (both humans and agents). Johnson et al [16] stress the importance of taking the interdependence of the team task into account when designing agents that are to function in a human-agent team. They promote a teamwork-centered approach when designing autonomous systems, called *coactive design*.

In this paper we will analyze literature on teamwork and explanation, and argue that explanation plays an important role in achieving good human-agent team performance. Namely, in order to coordinate actions, it is important that team members understand and predict each other's behavior, and explanations can help to improve insight in other team members' behavior. Consequently, we argue that to develop agents that perform well in human-agent teams, the agents should be equipped with explanation capabilities.

Furthermore, we will describe an empirical study investigating the role of explanation in human-agent teamwork. For the study, we will use the BlocksWorld for Teams (BW4T) testbed for team coordination [18]. In BW4T, a team of players has to perform a joint task in a virtual environment. The players are highly interdependent, and the performance of the team strongly depends on the level of coordination among the players.

The outline of this paper is as follows. In section 2, we will discuss literature on teamwork and explanation, and motivate why explanation is important in human-agent teamwork. In section 3, we will describe the BW4T coordination testbed for investigating teamwork. In section 4, we describe the experiment that examines the effect of explanations about agent behavior on the coordination in human-agent teams. In section 5, we end the paper with a conclusion.

2 Background

In this section we will discuss human teamwork, human-agent teamwork, and explanation in human-agent teams, respectively. Human teamwork has been studied for several decades. Compared to the large body of literature concerning human teamwork, there is relatively little literature on human-agent teamwork. The work on human-agent teamwork builds concepts and theories that were developed in research on human teamwork. Therefore, before we discuss human-agent teamwork specifically, we first provide a short introduction to human teamwork.

2.1 Human teamwork

There are two main streams in the literature on human teamwork. In the first, the concept of transactive memory is used to explain teamwork, and in the second, the concept of shared mental models is used to explain teamwork. We will describe both views.

Transactive memory systems The theory of transactive memory was first introduced by Wegner [35]. A transactive memory system (TMS) is a memory

system that is distributed across different team members. In a TMS, each of the team members has 1) knowledge that captures his or her own expertise, and 2) knowledge about who knows what. The knowledge that needs to be remembered is thus divided over the different team members. The assumption is that it is more efficient for an individual to remember who has knowledge on a certain topic than remembering all the details by oneself.

In order to use TMS theory for the explanation and prediction of team performance, different ways to measure TMS have been proposed [24, 1, 27]. Moreland et al [24], for example, distinguished three components of TSM: specialization, credibility and coordination. The specialization component refers to the level of knowledge differentiation within the team. Credibility refers to team members' beliefs about the accuracy of other members' knowledge. Coordination refers to team members' ability to work together efficiently.

Results of TMS as a determinant of performance are promising [1, 21, 25]. However, a real consensus among researchers on how to measure TMS is lacking. First, there is no commonly accepted theory on which components comprise TSM. Second, there are different ways to measure a team's performance on these components.

Shared mental models Mental models refer to the internal representations that humans have of the world around them. Mental models enable humans to understand, explain and predict of the systems in their environment [28]. In the context of teamwork, mental models can help individuals to understand the behavior of other team members and to predict their future actions. This allows the individuals to anticipate their own actions to the expected behavior of others. It is argued that in order to coordinate the actions of different team members well, it is important that the team members have similar mental models: shared mental models (SMM) [6]. Most researchers classify SMM into two broad dimensions: task-related knowledge and team-related knowledge (e.g. [7]). Task-related knowledge concerns knowledge about how to achieve the task, the current status of task achievement, etc. Team-related knowledge concerns knowledge and capabilities of other team members, what they are currently intending or doing, etc. Experimental results trying to demonstrate the effects of sharedness of mental models on team performance are promising. However, like for TMS, there is no common method for measuring the sharedness of mental models [23].

Relation between TMS and SMM There is little interaction between the research fields of TMS and SMM. An exception is the work of Nandkeolyar [25], who compared both theories on their predictive power on team learning and team effectiveness. He found that in most cases high levels of TMS components (specialization, coordination and credibility) and high levels SMM both predicted team performance well. However, in same cases, high levels of SMM did not result in high team performance, especially when teams scored high on TMS specialization and credibility.

Researchers from both sides have stated that one theory is an extention of the other. Shared knowledge in SMM theory can be seen as a team member's knowledge about who knows what in TMS theory. The other way around, a team member's knowledge about who knows what in TMS theory can be seen as shared knowledge in SMM theory. Whether both theories only provide a different vocabulary for the same processes or describe distinctive phenomena, the two theories have a different focus. TMS focuses more on the divideness of knowledge and SMM focuses more on the sharedness of knowledge. Both sides, however, do acknowledge that some of both is needed. Without any shared knowledge, it is not possible to coordinate actions, but totally overlapping knowledge leads a single minded view on tasks, also called *groupthink* [14].

2.2 Human-agent teamwork

Literature shows that sharedness and dividedness of knowledge are both important in human teamwork. In this section, we argue that sharedness and dividedness of knowledge are at least as important in human-agent teamwork.

Dividedness of knowledge is particularly important in human-agent teams because agents and humans have different strenghts and capabilities. For example, agents may be better at remembering a large amount of data than humans, but humans are often better at recognizing danger than agents. Both humans and agents even have capabilities that the other does not have. On the one hand, there is no human that can calculate as fast as an agent can, but on the other hand, there are no agents that can break the ice (socially). To fully benefit of the strenghts and capabilities that the members of a human-agent team offer, the tasks should be divided over the team members in such a way that each team member performs the tasks that best suit his or her capabilities and knowledge. For most tasks, especially the complex ones, this will lead to a division of knowledge over team members.

Sharedness of knowledge is important in human-agent teams to coordinate actions, especially because knowledge and capabilities are often divided over team members. When the members in a team have different strenghts, they must be are aware of each others' specialties in order to allocate subtasks to the right team member. Moreover, intially, humans know less about the behavior of an agent team member than the behavior of a human team member. Namely, being a human already reveals many properties of a team member, e.g. memory capacity, speed of doing tasks. Among agent team members, there is more diversity concerning these properties. Therefore, it is especially important that mental models about what team members know and can do are shared for the coordiantion of actions. In line with this argument, several approaches for team agents have been proposed that are explicitly based on SMM theory [19, 37].

In literature on human-machine interaction, there is a shift of attiontion from dividedness towards sharedness in human-agent teams. In the seventies, Sheridan and Verplank [29] introduced different levels of autonomy. At the highest autonomy level, the computer decides everything, acts autonomously, and ignores the human, and at the lowest autonomy level, the computer offers no assistance, and

the human must take all decisions and actions. This model of autonomy levels thus focuses on how tasks are divided over machines and humans. Johnson and colleagues [17] argue that the levels autonomy model falls short on the actual complexity of effective human-agent teamwork. They observe that humans and agents have different capabilities and argue that to combine their strengths, it is crucial to have good coordination in human-agent teams [3]. To coordinate actions it is necessary to exchange information about each others' goals, intentions, and observations. This stresses the importance of sharedness of knowledge.

We believe that explanation can contribute to coordination in human-agent teams in two ways. First, explanations about agent behavior can increase the sharedness of mental models by informing humans about the actions, observations and intentions of the agents. With this knowledge, humans will be able to better understand and predict new agent behavior, which will make it easier to coordinate actions. Second, explanations about agent behavior can increase humans' trust in agents. Members of a human-agent team usually have different knowledge and capabilities. So when a team member provides information other team members, e.g. informing about an intention, they will only use that information to coordinate their actions when they trust the team member. Having insight in another's reasoning increases trust, and in human-agents teams trust will improve coordination. In the next section, we will provide a short overview of research of explaining intelligent systems.

2.3 Explanation in human-agent teams

To discuss different applications in which intelligent system behavior is explained, we will use Sycara and Lewis' [31] distinction of different roles of software agents in human-agent teams. According to them, agents in a human-agent team can have the role of individual assistent, team assistent and equal team member. We will discuss the explanation of intelligent system behavior for each of these roles.

In the first role, an agent provides individual assistance to a human. In that case, the agent cooperates with only one human, who may or may not be part of a bigger team. Examples of systems providing explanations that can be seen as individual assistents are expert systems and recommender systems. These types of systems both support a single human user in making decisions. The explanation of intelligent system behavior was first researched in the field expert systems. It was discovered that to accept an advice or diagnose of an expert system, users want to know how and why a certain outcome was reached [30, 36, 11]. Aims of explanations in expert systems are increasing user acceptance, trust, ease of use, usefulness and user satisfaction [10]. Aims of of explanations in recommender systems are transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, satisfaction [32].

In the second role, an agent provides team assistance. A team assistant agent cooperates with all team members, usually to support coordination activities in the team [33]. The concept of team assistence is relatively young, and we are not aware of explanation approaches for team assistent agents.

In the third role, an agent acts as a (more or less) equal team member. In this role, an agent performs the reasoning and tasks of a human teammate. Virtual training is a field in which the behavior of agents in the role of an equal team member is explained. In virtual training, intelligent agents are used to play a trainee's colleagues, opponents or team members. Several approaches for explaining the behavior of such agent have been proposed [15, 34, 8, 12]. Explanations in virtual training aim to increase the trainee's understanding of the played session, and thereby support learning.

Different roles of agents in teams yield different types of explantions. Expert system behavior (where the system has the role of personal assistant), for example, is exlained by traces of rules that were applied and the justification behind those rules [11]. Behavior of agents in virtual training (where the agent has the role of equal team member) is explained in terms of goals and intentions [8, 12]. The differents between expert systems on the one hand, and virtual intelligent agents on the other hand, is that the behavior of the latter more closely resambles human behavior. Humans explain and understand their own and others' behavior in terms of the underlying mental concepts such as desires, plans, beliefs and intentions [22, 20]. In Dennett's words [9], people adopt the *intentional stance* towards virtual intelligent agents, i.e. they attribute beliefs and goals to them in order to understand their behavior. Thus, the role of the agent in a human-agent team should be taken into account when developing its explanation capabilities.

In the remainder of this paper we will discuss how to study the effects of explanation in human-agent teamwork. The BlocksWorld for Teams coordination testbed provides a mean to investigate human-agent teamwork. We will first describe the testbed itself, and subsequently, a study we performed in the testbed. Agents in BlocksWorld for Teams have the role of an equal team member.

3 The BW4T coordination testbed

BlocksWorld for Teams (BW4T) is a testbed for team coordination [18]. In the BW4T testbed, teams of humans, agents, or humans and agents can perform a task that requires coordination in a controlled environment. We therefore believe that the BW4T testbed is a useful tool for studying teamwork. The task is simple to learn and it is possible to manipulate all conditions in the environment, but at the same time, there are many interdependencies among the different players and complex process arise. In this section we will describe the BW4T task, discuss the behavior of a BW4T agent, and discuss the implementation of a BW4T agent.

3.1 The BW4T team task

The BW4T task can be performed by human-human, agent-agent and human-agent teams of variable sizes. The team goal is to jointly deliver a sequence of colored blocks in a particular order as fast as possible. A complicating factor

is that the players (human or agent) cannot see each other. Figure 3.1 displays a screenshot of a BW4T game session, showing the environment in which the players have to search for blocks. The left picture displays all blocks and players in the game, and the right picture shows what one player can see. A player can only see the blocks in a room when he is inside that room. The status bar below the Dropzone (gray area) shows which blocks need to be delivered.

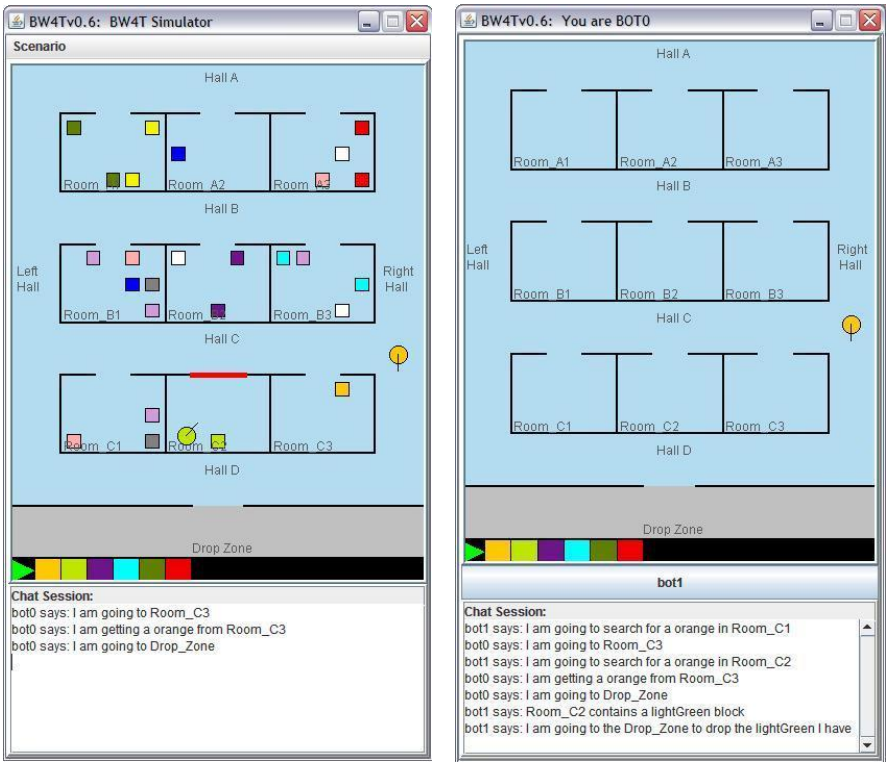

Fig. 1. Simulator view (left) and agent view (right). The blocks that need to be delivered are respectively orange, light green, dark purple, light blue, dark green and red. Bot0 in the right hall is holding an orange block and bot1 in room C2 is holding a light green block.

To deliver a block successfully, a player has to go to a block of the right color, pick it up and drop it in the Dropzone. A player can only carry one block at a time. When a player drops a block of the wrong color in the Dropzone or any block in a hall, the block disappears from the game. Human players can perform actions in the environment through a menu that appears on a right mouse-button click. The menu offers options to go to a place (room, hall or Dropzone), pick up a block, drop a block and send messages.

A team's performance on the BW4T task is measured by the speed of completing the task. BW4T is designed such that the task involves a large amount of interdependence among the players, and requires coordination to achieve a good performance. For instance, it is inefficient when one player is searching in a room that has just been checked by another. And if a player is going to deliver a particular block, the others should not do that as well. To coordinate, players can send messages to each other, which appear in the chatbox below the Dropzone. Players can inform others about what they do, where they are and what they see. Furthermore, players can see the same status bar. So when a player delivers a block of the right color, the other players will know. Finally, only one player can be inside a room or the Dropzone at the same time. When a player tries to enter a room that is occupied, a red bar appears indicating that someone is inside.

3.2 Behavior of a BW4T agent

Developing an agent that can perform the BW4T task on its own is rather straightforward. The agent needs to be able to search for blocks and deliver blocks, and it has to plan its behavior. Planning involves deciding what to do (search or deliver), where to search for blocks and which block to deliver. There are several strategies to perform the BW4T task. The agent can for instance search all needed blocks and then deliver them. It can also search for the next block in the sequence and deliver it once found, or keep checking rooms on the way to the Dropzone to deliver a block.

The agent's behavior gets more complex when there is a team of players involved. Each of the agent's has to coordinate its behavior with the others to avoid that a room is checked twice, or that two agents are delivering a block of the same color when only one block of that color is needed. To coordinate, the players have to update others about their activities and percepts, e.g. tell others what they are going to do and which blocks they found in which rooms. Moreover, they have to adapt their own behavior to messages they receive from others. For example, if a red block needs to be delivered and another player says it is going to deliver that block, it is better to search for the next block in the sequence.

When the behavior of the other players in a team is known, it is sufficient to send updates and process updates from others for effective task performance. However, in applications with human-agent teams, usually the behavior of the others is not completely known. The behavior of the agents may be designed by different developers, and behavior of human players can never be completely predicted as humans tend to vary their strategy, make mistakes and forget things. It may happen, for instance, that a player tells that there is a yellow block in room C1, but once you arrive it is not there, or that a player announces that he is going to deliver an orange block, but actually does not, or that someone delivered a white block, even though you had told to deliver it. Therefore, a BW4T agent should be able to deal with unexpected events.

3.3 Implementation of a BW4T agent

Currently, there are two ways to implement a BW4T agent. The first way is to use Java. BW4T is implemented in Java and offers a basic agent class in which the behavior of a BW4T agent can be specified. The second way is to use `GOAL` [13], a BDI-based (Belief Desire Intention) programming language.

BDI-based programming languages offer the possibility to represent an agent's behavior in terms of its beliefs and goals, and a BDI agent's actions are determined by a deliberation process on its beliefs and goals. The BDI-based agent programming paradigm is based on Bratman's theory of human practical reasoning, in which human reasoning is described with the notions of belief, desire and intention [5]. Rao and Georgeff developed a BDI-based software model [26] based on Bratman's theory. A typical BDI deliberation cycle contains the following steps: (i) perceive 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 subgoal.

Currently, there is a set of BDI-based agent programming languages [2] and `GOAL` is one of them. A connection has been established between BW4T and `GOAL`, which makes it possible to implement BW4T agents in `GOAL`.

4 Experiment

In this section we describe the experiment performed in BW4T. As motivated in section 2, we believe that human-agent teams in which agents explain their behavior coordinate better than human-agent teams in which agents do not explain their behavior. In the experiment, we will use performance on the BW4T task to measure the level of coordination in human-agent teams. Our hypothesis is that human-agent teams in which agents explain their behavior perform better on the BW4T task than human-agent teams in which agents do not explain their behavior.

4.1 Method

Design. The experiment has a within-subjects design with an explanation and a no-explanation condition. In the explanation condition, the subjects cooperate with an agent explaining its behavior, and in the no-explanation condition, subjects cooperate with an agent that does not explain its behavior. The order of the two conditions, explanation and no-explanation were assigned counterbalanced to the subjects, to correct for possible learning effects from the first to the second trial.

Subjects. A total of 16 subjects (male = 14, female = 2) with an average age of 27 (sd=3.5) participated in the experiments.

Materials. We used the BW4T testbed described in section 3. In order to investigate the effects of an agent's explanation on human-agent team performance, we developed a BW4T agent that is able to explain its behavior. We implemented the agent in GOAL.

The agent's behavior is formed by the following rules. The agent starts to check rooms and once it knows about a block that can be delivered, it starts to deliver that block. The agent uses information about blocks in rooms received from other players. When another player announces that he is going to check a particular room, the agent will not check that room. When another player tells that he is going to deliver a block, the agent will start to search or deliver the next block in the sequence. The agent is able to deal with humans that vary their strategy, make mistakes and forget to tell things. Namely, the agent revises its plans when a room contains other blocks than it expected, and when the agent holds a block that is not needed anymore, it will drop the block in a room. Thus, in general, the agent is cooperative and assumes that other players are cooperative as well.

The following GOAL code shows a part of the agent's planning behavior in which it decides to either deliver a block or check a room.

```
IF a-goal(deliverSequence), bel(me(Me),available(Me),
   toPickUp(Block,Color),in(Block,Room))
THEN adopt(delivered(Block)) + insert(delivering(Me,Block)).

IF a-goal(deliverSequence), bel(me(Me),available(Me),
   not(toPickUp(Block,Color)),nextRoomInSeq(Room),
   not(checked(Room)),not(checking(_,Room)))
THEN adopt(checked(Room)) + insert(checking(Me,Room)).
```

The first if-then rule states that if it is the agent's goal to deliver the sequence of blocks, and it believes that it is available to do something and that there is a block that can be picked up, then it will adopt the goal to deliver that block and obtains the belief that it is delivering that block. the second if-then rule state that if the agent if it is the agent's goal to deliver the sequence of blocks, and it believes that it is available to do something, there is no block that can be picked up, and the next room that has not been checked is not already being checked by someone else, then the agent will adopt the goal to check that room and obtains the belief that it is checking that room.

As the aim of this study is to study the effects of explanations about agent behavior on coordination in human-agent teams, the agent needs to be able to explain its behavior. In section 2, we argued that agents that play the role of an equal team member are considered intentional. In other words, we understand their behavior by attributing beliefs, goals and intentions to them. We therefore believe that the beliefs, goals and intentions underlying the agent's actions comprise useful explanations about its behavior. The implementation in GOAL allowed us to explain the agent's behavior in terms of beliefs, goals and intentions [12].

To explore the effect of explaining agent behavior on coordination in human-agent teams, we need to be able to manipulate the agent's communication behavior. Inspired on the KaOS policy framework [4], we use policies to regulate the agent's communication behavior, so we do not have to change the agent's programming code. We distinguish the following three communication policies.

1. Inform other players about your observations
2. Inform other players about your actions
3. Provide explanations for your actions

The first policy entails that if the agent observes something in the virtual environment, he sends a message to inform all other players about his observation. Such messages are, for example, 'Room A1 contains a pink block and a dark blue block' and 'Room B2 is empty'. The second policy prescribes that if the agent performs an action, he has to send a message to inform all other players about it. Messages informing about actions are for instance 'Im going to Room C1', 'I picked up a red block' and 'I just dropped a gray block'. The third policy prescribes the agent to explain an action, that is, to provide the underlying goal of that action. In the next section we will discuss the explanation of actions in more detail. Examples explanations for actions are 'I am going to Room B3 to search for an orange block' and 'I am going to Room C2 to deliver a light green block'.

In the explanation condition, the agent adhered to all three communication policies, and in the no-explanation condition, only communication policies 1 and 2 were applied. Thus, the agent equally often provided updates in both conditions, but the updates in the explanation condition were longer than those in the no-explanation condition.

Procedure. The subjects received an explanation of the BW4T task and how to direct their 'bot'. Subsequently, they had to play a training session, in which they had to deliver three blocks on their own. The training session was included to make sure that the subjects completely understood the game, and to give them time to think about their strategy in the actual trials. No agent participated in the training session yet, to prevent that it would shape the subjects' expectations about the agents in the trial sessions.

For the two trial sessions, subjects were instructed to perform the task with the agent as a team, as fast as possible. They were told that the agent could show any kind of behavior, e.g. not search in the right places or not take the subject's messages into account, but that the agent would not lie to them. In both trial sessions, the human-agent team delivered six blocks of different colors. The colors and positions of the blocks differed per session, but the total traveling distance to deliver all blocks was the same. The order of the two conditions, explanation and no-explanation were assigned counter-balanced to the subjects, to correct for possible learning effects from the first to the second trial. After both sessions, the subjects were asked to fill in a short questionnaire.

4.2 Results

The time of completing the BW4T task was used as a measure for team performance. In the explanation condition, the average time (n=16) to complete the task was 596 seconds (sd=118), and in the no-explanation condition the average time was 593 seconds (sd=81). These averages are obviously not significant (paired t-test: p=0.95).

We also examined if there was a learning effect between the first and second session. The average time (n=16) to complete the sessions was 617 seconds (sd=118) for the first session, and 572 seconds (sd=76) for the second session. The results show that the subjects completed the task faster in the second session than in the first session, but the difference is not significant (paired t-test: p=0.26).

In the questionnaire administered after each session, we asked subjects to judge their own, the agent's and their common performance on a scale from 1 to 7. Table 1 shows the averages in both the explanation condition (EX) and the no-explanation condition (NE).

	EX	NE
I was effectively performing the task	5.9 (sd=0.7)	5.8 (sd=1.1)
The agent was effectively performing the task	6.0 (sd=1.3)	5.5 (sd=1.3)
We were effectively performing the task as a team	5.7 (sd=1.6)	5.1 (sd=1.7)

Table 1. Average estimation of performance on a 1-7 scale (n=16).

The results are not significant (paired t-tests: p=0.67, p=0.36, p=0.41, respectively), but for all questions and in particular for agent and team performance, the subjects judged performance on average higher in the explanation condition than in the no-explanation condition, even though no actual differences in performance were found.

In order to investigate how well subjects evaluate performance, we calculated the correlations between the self-evaluations in Table 1 and the actual team performances. Surprisingly, the subjects' self-evaluations have a low or even negative correlation with the actual performances. Three of the negative correlations are significant (α=0.05): evaluated human performance and actual team performance in the no-explanation condition (R=-0.49), evaluated agent performance and actual team performance in the explanation condition (R=-0.50), and evaluated team performance and actual team performance in the explanation condition (R=-0.55). The results show that subjects make better estimates of their own performance in the explanation condition, and better estimates of the agent's and the team's performance in the no-explanation condition.

In the questionnaire, we also asked the subjects to judge how well they understood the actions and motivations of the agents, and how well the agents seemed

to understand their actions and motivations. The results in Table 2 show that the subjects had a significantly better idea of what the agent was doing in the explanation condition than in the no-explanation condition (paired t-test: p=0.030). Though the other results are not significant, for all questions understanding was on average rated higher in the explanation than in the no-explanation condition (paired t-test: p=0.74, p=0.65, p=0.47, respectively).

	EX	NE
I had a good idea of what the agent was doing	6.1 (sd=1.0)	5.1 (sd=1.4)
The agent seemed to have a good idea of what I was doing	5.8 (sd=1.1)	5.7 (sd=1.0)
I understood the reasons behind the agent's behavior	5.9 (sd=1.2)	5.7 (sd=1.5)
The agent seemed to understand the reasons behind my behavior	5.6 (sd=1.0)	5.3 (sd=1.9)

Table 2. Average understanding of behavior on a 1-7 scale (n=16).

Finally, we asked subjects if the agent provided *too little*, *just enough*, or *too much* information. In the explanation condition, 1 subject thought that the agents provided too little information, and all other 15 subjects thought that the agent provided just enough information. A chi-square goodness of fit test shows that the result is significant (χ^2=26.4, p<0.001). In the no-explanation condition, 10 subjects indicated that the agents provided too little information, while 6 subjects indicated that the provided information was just enough. This result is significant as well (χ^2=9.5, p=0.009). Thus, in general subjects preferred the amount of information in the explanation condition over the amount of information in the no-explanation condition.

4.3 Discussion

We found no significant differences between human-agent team performance in the explanation and the no-explanation condition. Therefore, the results do not support our hypothesis that explanations about agent behavior improve human-agent team performance on the BW4T task. The experience of the subjects, however, was affected by the agent's explanations. The subjects' ratings of their idea of what the agent was doing was significantly higher in the explanation condition than in the no-explanation condition. Furthermore, a significant number of subjects believed that the agent in the no-explanation condition provided too little information, whereas a significant number of subjects indicated that the agent in the explanation condition provided just enough information.

With a larger number of subjects, more of the results obtained from the questionnaire may have been significant. Namely, all of the subjects' ratings

are higher for the explanation condition than for the no-explanation condition, both concerning self-evaluations on performance as understanding of each other's actions. It is not probable that the difference in performance on both conditions quickly would have become significant with a larger number of subjects, since the performances on both conditions are rather similar.

There are several possible explanations for the similar team performances on both conditions. We provide five of them. First, subjects may have lost time in processing the agent's explanations, which then was compensated by a more efficient task completion. The robots in BW4T move slowly on purpose to provide players sufficient time to communicate, and think and process information. However, at some points in the game many actions have to be done at once (enter a room, go to a block, pick up a block, go to the Dropzone, and communicate about your actions) despite of the slow speed of the robots. Thus, at those time points, processing explanations may lead to time loss.

Second, the subjects may have anticipated a cooperative agent. Though we told them that the agent could perform any behavior and made them aware of possible strategies, several of the subjects reported that their strategy was to behave as if the agent was cooperative until they would find out otherwise. With such a strategy, explanations do not contribute to a quicker adaptation to the agent's behavior as the subject's initial behavior already makes the right assumptions about the agent's behavior. It would be interesting to conduct an experiment with a less cooperative or capable agent, e.g. one that cannot process certain messages or is colorblind, to see if explanations help subjects to quicker adapt to the gaps in the agent's capabilities.

Third, the task may involve too much noise. Some of the subjects, for instance, reported that they mistook one color for another (e.g. yellow and light green), which caused a serious delay. Other subjects said that they changed their strategy after the first trial, e.g. they let the agent deliver all blocks. Furthermore, though the blocks are evenly spread over the rooms in different trials, there is a luck factor involved in finding blocks. This factor can be decreased by letting the team deliver more blocks, but adding blocks also gives the subjects more time to learn the agent's behavior, which decreases the expected effect of providing explanations. In conclusion, noise factors like these may have wiped out the effects of explanation on team performance.

Fourth, the task may be too simple to show an effect. In most situations, the rationale behind the agent's behavior can be deduced from its actions.

Finally, the agent always explained its actions by the goals they aimed to achieve. The advantage of such explanations is that they are immediately derivable from the mental state of a BDI agent. Possibly, when extending the agent's explanation capabilities, e.g. by adding information about the agent's strategies, the explanations would become more useful and have a bigger effect on team performance.

5 Conclusion

In this paper, we discussed literature on human teamwork, human-agent teamwork and the explanation of intelligent systems and agents. We argued that explanation of the behavior of agents in a human-agent teams can contribute to team performance in two ways. First, when team members have more shared knowledge, e.g. about their current activities and plans, it is easier to coordinate their actions. Second, explanations can increase trust in a team member, which also facilitates the coordination of actions.

Furthermore, we presented a study in the BW4T coordination testbed that examined the effects of agents explaining their behavior on coordination in human-agent teams. A first result was that, against our expectations, explanations about agent behavior did not lead to better team performance. In the discussion we suggested several explanations for these results, e.g. the task being too simple. A second result was that, in correspondence to our expectations, humans indicated that they better understood the agent's behavior when they received explanations about it.

Though the BW4T task is simple, we believe that it offers a good platform for investigating human-agent teamwork. In order to learn more from experiments with the BW4T testbed, we will use more diverse conditions in future research than we used in the study described in this paper. For instance, the effects of no communication at all and an overload of explanations could be compared to the current results. Furthermore, in future work we want to measure more dependent variables in the experiments. Besides time of completing the task, we will also measure the sharedness of knowledge between team members, and the trust humans have in agents. Such research will give insight in whether concepts that were adopted from literature on human teamwork also apply to human-agent teamwork.

Acknowledgments

This research has been supported by the EOARD, grant nr. 103015.

References

1. J. Austin. Transactive memory in organizational groups: the effects of content, consensus, specialization and accuracy on group performance. *Journal of Applied Psychology*, 88(5):866–878, 2003.
2. R. H. Bordini, M. Dastani, and A. E. F. Seghrouchni, editors. *Multi-Agent Programming: Languages, Tools and Applications*. Springer, 2009.
3. J. M. Bradshaw, P. Feltovich, and M. Johnson. *Handbook of Human-Machine Interaction*, chapter Human-Agent Interaction, page 283302. Ashgate, 2011.
4. J. Bradshaw et al. Representation and reasoning about daml-based policy and domain services in KAoS. In *Proceedings of AAMAS03*. ACM Press, 2003.
5. M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, Massachusets, 1987.

6. J. A. Cannon-Bowers, E. Salas, and S. Converse. Shared mental models in expert team decision making. *Individual and group decision making Current issues*, 39(3-4):221–246, 1993.
7. N. J. Cooke, E. Salas, J. A. Cannon-Bowers, and R. J. Stout. Measuring team knowledge. *Human Factors*, 42(1):151–173, 2000.
8. 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.
9. D. Dennett. *The Intentional Stance*. MIT Press, 1987.
10. J. Dhaliwal and I. Benbasat. The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. *Information systems research*, 7(6):243–361, 1996.
11. S. Gregor and I. Benbasat. Explanation from intelligent systems: theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530, 1999.
12. M. Harbers, K. Van den Bosch, and J.-J. Meyer. Design and evaluation of explainable agents. Proceedings of IAT10, 2010.
13. K. Hindriks. *Multi-Agent Programming: Languages, Tools and Applications*, chapter Programming Rational Agents in GOAL, pages 119–157. Springer, 2009.
14. I. L. Janis. *Victims of Groupthink*. Houghton Mifflin, Boston, 1972.
15. L. Johnson. Agents that learn to explain themselves. In *Proc. of the 12th Nat. Conf. on Artificial Intelligence*, pages 1257–1263, 1994.
16. M. Johnson, J. Bradshaw, P. Feltovich, C. Jonker, M. Van Riemsdijk, and M. Sierhuis. Coactive design: Why interdependence must shape autonomy. In *Coordination, Organizations, Institutions, and Norms in Agent Systems*, in press.
17. M. Johnson, J. M. Bradshaw, P. J. Feltovich, R. R. Hoffman, C. Jonker, B. van Riemsdijk, and M. Sierhuis. Beyond cooperative robotics: The central role of interdependence in coactive design. *IEEE Intelligent Systems*, 26:81–88, 2011.
18. M. Johnson, C. Jonker, M. Van Riemsdijk, P. Feltovich, and J. Bradshaw. Joint activity testbed: Blocks world for teams (BW4T). In *Proceedings of ESAW09*, pages 254–256. Springer, 2009.
19. C. M. Jonker, M. B. van Riemsdijk, and B. Vermeulen. Shared mental models - a conceptual analysis. In *Proceedings of COIN@AAMASMALLOW'2010*, pages 132–151, 2010.
20. F. Keil. Explanation and understanding. *Annual Reviews Psychology*, 57:227–254, 2006.
21. K. Lewis. Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology*, 88(4):587–604, 2003.
22. B. Malle. How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1):23–48, 1999.
23. S. Mohammed, R. Klimoski, and J. R. Rentsch. The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, 3(2):123–165, 2000.
24. R. L. Moreland and L. Myaskovsky. Exploring the performance benefits of group training: Transactive memory or improved communication? *Organizational Behavior and Human Decision Processes*, 82(1):117–133, 2000.
25. A. K. Nandkeolyar. *How do teams learn? shared mental models and transactive memory systems as determinants of team learning and effectiveness*. PhD thesis, University of Iowa, 2008.
26. A. Rao and M. Georgeff. BDI-agents: From theory to practice. In *Proceedings of ICMAS'95*, 1995.

27. D. Rau. Top management team transactive memory, information gathering, and perceptual accuracy. *Journal of Business Research*, 59(4):416–424, 2006.
28. W. Rouse and N. M. Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363, 1984.
29. T. B. Sheridan and W. L. Verplank. Human and computer control of undersea teleoperators (tech. rep.). Technical report, ManMachine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1978.
30. W. Swartout and J. Moore. *Second-Generation Expert Systems*, chapter Explanation in Second-Generation Expert Systems, pages 543–585. Springer-Verlag, New York, 1993.
31. K. Sycara and M. Lewis. Integrating intelligent agents into human teams. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, 2002.
32. N. Tintarev and J. Masthoff. A survey of explanations in recommender systems. In *Proceeding of the International Conference on Data Engineering Workshop*, Washington, DC, 2007. IEEE Computer Society.
33. J. van Diggelen, R.-J. Beun, and P. J. Werkhoven. Intelligent assistants in crisis management: from pda to tda. In *Proceedings of BNAIC 2009*, 2009.
34. 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.
35. D. M. Wegner. Transactive memory: A contemporary analysis of the group mind. *Theories of group behavior, edited by B. Mullen and G. R. Goethals*, pages 185–208, 1987.
36. R. Ye and P. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2):157–172, 1995.
37. J. Yen, X. Fan, S. Sun, T. Hanratty, and J. Dumer. Agents with shared mental models for enhancing team decision-makings. *Deision Support Systems*, 41:634–653, 2006.