AFRL-AFOSR-UK-TR-2011-0046

Self-explaining Agents
A Study in the BW4T Testbed for Team Coordination

Maaike Harbers
Jeffrey Bradshaw
Matthew Johnson

TNO Defense, Security and Safety
Human Factors
Kampweg 5
Soesterberg, The Netherlands 3769 DE

Paul Feltovich
Karel Van den Bosch
John-Jules Meyer

Florida Institute for Human and Machine Cognition
40 South Alcaniz Street
Pensacola, FL 32502

EOARD GRANT 10-3015

October 2011

Final Report for 15 June 2010 to 15 June 2011

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
11-10-2011	Final	14 June 2010 - 14 June 2011

4. TITLE AND SUBTITLE
Self-explaining Agents

A Study in the BW4T Testbed for Team Coordination

5a. CONTRACT NUMBER
FA8655-10-1-3015

5b. GRANT NUMBER
10-3015

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Harbers, Maaike; Bradshaw, Jeffrey, M.; Johnson, Matthew;

Feltovich, Paul; Van den Bosch, Karel; Meyer, John-Jules

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO Defense, Security and Safety
Human Factors
Kampweg 5
Soesterberg, The Netherlands 3769 DE

Florida Institute for Human and Machine Cognition
40 South Alcaniz Street
Pensacola, FL 32502

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR/RSW (EOARD)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2011-0046

12. DISTRIBUTION / AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. 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. Coordination in human-agent teams can be improved by giving humans insight in the behavior of the agents. When humans are able to understand and predict an agent's behavior, they can more easily adapt their own behavior to that of the agent. One way to achieve such understanding is by letting agents explain their behavior. This report presents a study in the BW4T coordination testbed that examines the effects of agents explaining their behavior on coordination in human-agent teams. The results 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.

15. SUBJECT TERMS
Explanation, Coordination, Human-Agent Teams

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			JAMES LAWTON Ph. D.
UNCLAS	UNCLAS	UNCLAS	SAR	7	**19b. TELEPHONE NUMBER** *(include area code)* +44 (0)1895 616187

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

INTRODUCTION

Humans easily adapt their behavior to entities with other cognitive abilities than their own. For instance, most people automatically choose simpler words when they talk to children than when they talk to adults, and many people are able to interact well with their pets even though those cannot speak at all. Also in human-agent interaction, the key of good interaction is not that the other, the agent in this case, acts as human-like as possible. Instead, agents should facilitate humans' adaptation to them by being transparent and observable [1]. Knowing more about an agent, e.g. its capabilities, goals and intentions, allows humans to better understand and predict the agent's behavior, and more easily adapt their own behavior to the agent. Therefore, explanations about agent behavior can improve coordination in human-agent teams.

Johnson et al [6] promote a teamwork-centered approach when designing autonomous systems, called coactive design. According to the coactive design approach, the design of autonomous agents should be led by the underlying interdependence of the joint activity in the human-agent system. In other words, understanding of the joint activities that the agents will undertake should be used to shape the implementation of agent capabilities. When human-agent teams have to perform tasks that involve a lot of coordination, it is important that the human(s) can understand and predict the behavior of the agent(s). Therefore, in such situations, agents should be capable of providing explanations that increase human understanding in their behavior.

Humans explain and understand their own and others' behavior in terms of the underlying mental concepts such as desires, plans, beliefs and intentions [9, 8]. In other words, they adopt the intentional stance towards others, i.e. they attribute beliefs and goals to them in order to understand their behavior [3]. Following this research, we believe that to support coordination in human-agent teams, agents can give humans insight in their behavior by revealing the underlying goals, beliefs and intentions.

Harbers et al [4] developed an approach for explainable agents in which agents explain their actions in terms of beliefs and goals. The approach was developed for virtual training systems, where virtual agents playing the trainee's team members explain their behavior to increase the trainee's understanding of the played session. The explainable agents are implemented in BDI-based (Belief Desire Intention) agent programming languages. That means that an agent's behavior is represented by beliefs, goals, plans and intentions [10], and its behavior is determined by a deliberation process on its mental concepts. The mental concepts underlying an action are also used to explain that action.

In this report, we describe an experiment that aims to show that BDI-based explanations about agent behavior improve coordination in human-agent teams. For that, we will use the BlocksWorld for Teams (BW4T) testbed for team coordination [7]. In BW4T, a team of players has to perform a joint task in a virtual environment, and the performance of the team strongly depends on the level of coordination among the players. Our hypothesis is that human-agent teams perform better on the BW4T task when agents explain their behavior than when they do not explain their behavior.

The outline of this report is as follows. First, we will describe the BW4T coordination testbed and discuss the development of a BW4T agent. Subsequently, we describe an experiment that examines the effect of explanations about agent behavior on the coordination in human-agent teams. Finally, we end the report with a conclusion.

THE BW4T COORDINATION TESTBED

BlocksWorld for Teams (BW4T) is a testbed for team coordination [7]. BW4T allows for games with human-human, agent-agent and human-agent teams of variable sizes. The 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 cannot see each other. Figure 1 shows the environment in which the players have to search for blocks. The left picture gives an overview of the game, and the right picture shows what a single player (human or agent) can see. The blocks are hidden in the rooms and only become visible when a player is inside a room. The status bar below the Dropzone (gray area) indicates which blocks need to be delivered.

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, pick up a block, drop a block and send messages. To successfully deliver a block, a player has to drop a block of the right color 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.

In order to effectively solve this task as a team, players have to coordinate their actions. For instance, it is inefficient when a player is searching in a room that has just been checked by another player. And if one 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 all others or one particular agent about what they do, where they are and what the see. Furthermore, all players can see the same status bar. So when a player delivers a block of the right color, all 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 which is occupied, a red bar appears indicating that someone is inside.

Performance on the BW4T task is measured by the speed of completing the task. BW4T is designed such that the task involves a lot of interdependence among different players, and requires effective coordination to achieve a good performance.

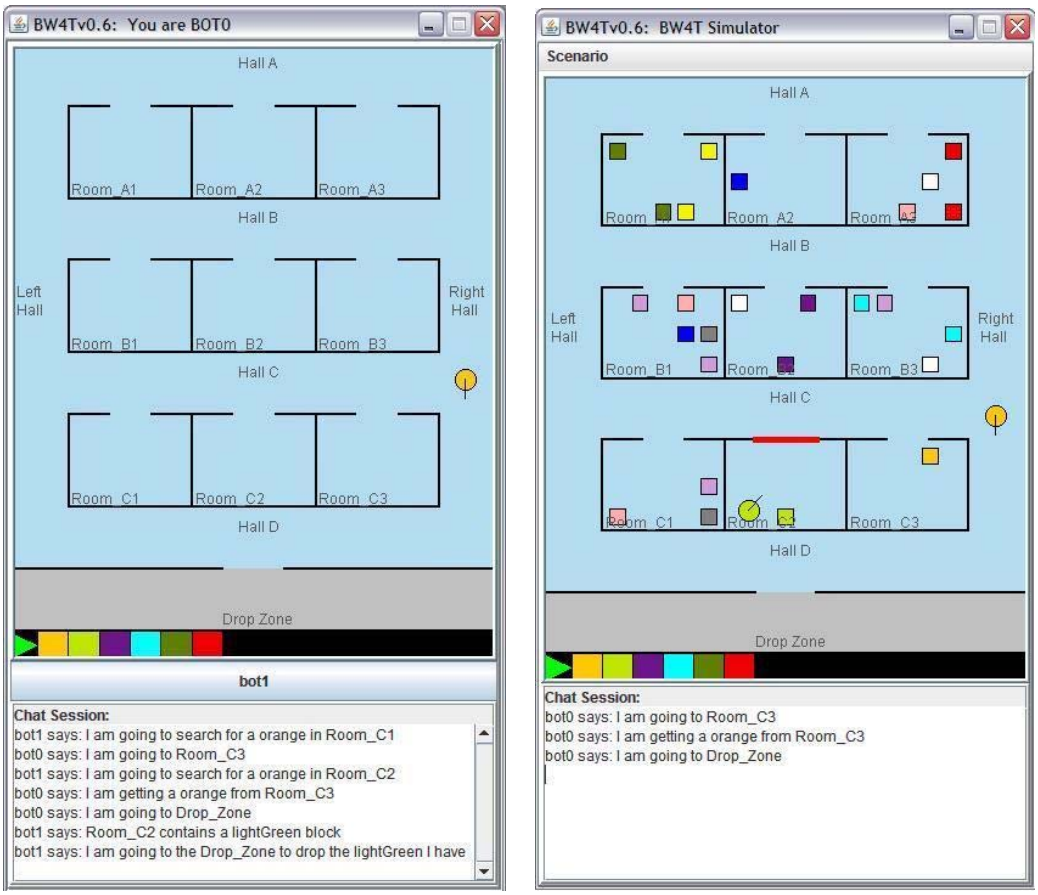

Figure 1: Simulator view and agent view

DEVELOPING A BW4T AGENT

In this section we discuss the development of a BW4T agent. We first discuss different strategies to perform the task. Subsequently, we discuss different ways in which the agent can communicate to other BW4T players. Finally, we discuss the agent's implementation.

Strategies to perform the BW4T task

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.

For our study, we designed a cooperative agent, which assumes that other players are cooperative as well. Its 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.

Communication policies

In this report, we aim to explore the effect of explaining agent behavior on coordination in human-agent teams, and therefore the agent's communication behavior needs to be changed. Inspired on the KaOS policy framework [2], we use policies to regulate the agent's 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 `I'm 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'.

Implementation

BW4T is implemented in Java and offers a basic agent class in which a BW4T agent's behavior can be implemented. At the TU Delft, a connection between BW4T and the BDI-based (Belief Desire Intention) programming language GOAL [5] has been established, which allows for the implementation of BW4T agents in GOAL. We have implemented our BW4T agent in GOAL because that facilitates the explanation of the agent's behavior.

GOAL offers the possibility to represent an agent's behavior in terms of its beliefs and goals. Action rules are used to specify which actions to select, given an agent's beliefs and goals. Besides beliefs, goals and action rules, a GOAL agent program also contains percept rules and an action specification section. A deliberation process on the agent's mental state ensures that percepts are processed, and that actions are selected and executed. A distinguishing feature of GOAL is its declarative goals, which describe what an agent wants to achieve, not how to achieve it. GOAL agents are committed to their goals and only remove a goal when it has been completely achieved.

To investigate the effect of explanations about agent behavior on coordination in human-agent teams, we had to make the BW4T agent implemented in GOAL able to explain its own behavior. Following Harbers et al's approach [4], we made the agent able to explain its behavior in terms of the goals and beliefs underlying its actions.

EXPERIMENT

In this section we describe the experiment performed in BW4T. As motivated in the introduction, 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. We expect 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.

Method

Design. The experiment has a within-subject design with an explanation and a no-explanation condition. In the explanation condition, the subjects have to cooperate with an agent explaining its behavior, and the no-explanation condition, the subjects have to cooperate with an agent that does not explain its behavior.
Subjects. 16 subjects (m=14,f=2) with an average age of 27 participated in the experiments.
Materials. We used the BW4T testbed and the agent described in the previous section. 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.

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 almost the same, and 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. This difference is not significant (paired t-test: p=0.26). However, the results indicate a trend towards faster completion of the task in the second than in the first session.

In the questionnaire registered 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	5.8
The agent was effectively performing the task	6.0	5.5
We were effectively performing the task as a team	5.7	5.1

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.

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 (alfa=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).

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	5.1
The agent seemed to have a good idea of what I was doing	5.8	5.7
I understood the reasons behind the agent’s behavior	5.9	5.7
The agent seemed to understand the reasons behind my behavior	5.6	5.3

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). n 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).

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. First, subjects may have lost time in processing the agent's explanations, which then was compensated by a more efficient task completion. This is not probable, since the robots in BW4T purposely move rather slow to provide players sufficient time to think.

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.

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.

CONCLUSION

In this report, we presented a study in the BW4T coordination testbed that examined the effects of agents explaining their behavior on coordination in human-agent teams. The results showed that explanations about agent behavior do not always lead to better team performance, but they do impact the user experience in a positive way. In the discussion, we suggested several possible explanations for the similar performances in both conditions, which could be further explored. Furthermore, BW4T allows for different team compositions, which offers another interesting direction for future research.

REFERENCES

[1] J. Bradshaw, P. Feltovich, and M. Johnson. Human-agent interaction. In G. Boy, editor, Handbook of Human-Computer Interaction, in press.

[2] J. Bradshaw et al. Representation and reasoning about daml-based policy and domain services in KAoS. In Proceedings of AAMAS03. ACM Press, 2003.

[3] D. Dennett. The Intentional Stance. MIT Press, 1987.

[4] M. Harbers, K. Van den Bosch, and J.-J. Meyer. Design and evaluation of explainable agents. Proceedings of IAT10, 2010.

[5] K. Hindriks. Multi-Agent Programming: Languages, Tools and Applications, chapter Programming Rational Agents in GOAL, pages 119-157, Springer, 2009.

[6] 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.

[7] 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.

[8] F. Keil. Explanation and understanding. Annual Reviews Psychology, 57:227-254, 2006.

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

[10] A. Rao and M. Georgeff. BDI-agents: From theory to practice. In Proceedings of ICMAS'95, 1995.