Delft University of Technology, Bachelor Seminar of Computer Science and Engineering

Influence of shared information on predictability in human-agent teams

Timea Nagy Ruben Verhagen Carolina Jorge
Dr. Myrthe Tielman
TU Delft

June 27, 2021

Abstract

Understanding trust in human-agent teams is of utmost importance if we want to ensure an efficient and effective collaboration. It is well known that predictability is a core component of trust, however it is still unclear what kind of information an agent should share in order to be perceived as predictable. Here we show that in a simple world setup with a noncomplicated task, there is no significant difference in the measured predictability between agents sharing information pertaining to only world knowledge, actions, world knowledge and actions or world knowledge, actions and explanations. However, previous experience with the framework used or having a technical background do greatly impact the perceived predictability. The small sample size and the data not being representative lead us to conclude that the study should be repeated with a larger and more diverse group of participants and a more complex world setup.

1 Introduction

Artificial intelligence and intelligent systems becoming fully autonomous is not the ultimate goal anymore. There is a steady increase in applications requiring that agents interact and collaborate with humans in progressively more complex scenarios. Differences between humans and AI are obvious, yet they can complement each other in multiple ways: people can provide context to the agent and keep its model accurate with the world; in turn the agent can "computationally instantiate their models of the world" and keep us updated with ongoing events [1]. To leverage this diversity in skills and abilities, more and more human-AI teams are emerging [2]. However, in order for such teams to be successful, it is crucial to understand the elements of teamwork and the ways both humans and agents behave in such configurations.

In their 2019 article "No AI is an island", Johnson and Vera argue that a key component of effective teamwork is teaming intelligence, which revolves around managing interdependence between the human and AI team members, claiming that "AI will only become an effective team player if it has an understanding of interdependence and is designed to support management of interdependencies with people" [3]. To understand how such interdependencies can shape the collaboration between humans and agents and how they can be translated

into concrete agent designs, the Coactive design framework developed by Johnson et al can be utilized. It identifies three requirements introduced by interdependencies in teamwork: observability, directability and predictability [4]. They further argue that "predictability is also essential to many teamwork patterns". It can also be seen from Stubbs et al's 2007 field study of Human Robot Interaction that these OPD requirements play a crucial role in team effectiveness [5]. Johnson and Bradshaw have greatly encapsulated the relevance of predictability with regards to trust and thus team effectiveness: "Predictability has been a long-standing cornerstone of trust" [6].

Moreover, the field of explainable AI is preoccupied with researching how making the agents more explainable can lead to achieving these requirements and thus increase the efficiency and effectiveness of human-agent teamwork. There are various works analysing how trust emerges between the human and agent teammates and how the level of explainability of the agent can influence this. For example, Verhagen et al claim that an understandable system enables predictability, thus bringing the pursuit of system understandability to the forefront of the development of agents part of human-agent teams [2]. This further confirms the necessity of investigating what kind of information an agent should share in order to be predictable to the human teammate, since sharing this relevant information leads to explanations that can "help users to increase confidence and trust" and that are considered crucial in collaborative tasks [7].

Therefore, the question arises: "what information should an agent share to be predictable to a human?", with predictability being defined as "the extent to which human users can estimate future or other functional system elements" [2]. By tackling this question, it is expected that valuable insight will be gained into what an agent should share with its human teammates in order to be perceived as predictable and thus enhance the efficiency and effectiveness of such collaborations. More concrete contributions of this paper are determining which type of information shared contributes to the highest predictability (measured both objectively and subjectively), seeing how this influences the efficiency of the team, whilst also analyzing the relationship between previous experience with such an agent and predictability. We will also tackle the question of how different types of explanations influence predictability.

The approach taken is that of conducting a controlled experiment within which participants are teamed up with agents sharing various levels of information to solve a given task within the Block Worlds for Teams (BW4T) test bed. The shared information is categorized into world knowledge and actions, with one agent sharing additional explanations relating to its actions.

This paper is outlined as follows. Section 2 clarifies the necessary theoretical concepts and is followed by the description of the BW4T test bed and task in Section 3. Section 4 lays out the experimental setup and the found results. The interpretation of the results, together with a discussion about responsible research, potential limitations and future work ensues in Section 5. The paper is concluded in Section 6.

2 Background

The previous section made it apparent that in order for human-agent teams to operate effectively and efficiently, mutual trust is paramount. Research points out that an important element of this trust is predictability, which brings forth the topic of understandable agents. Thus one must first gain insight into what makes an agent understandable.

In their 2021 work, Verhagen et al present a two dimensional framework that classifies agents into the following categories: incomprehensible, interpretable and understandable. They argue that transparency leads to interpretability and that understandability can be achieved through explainability. Furthermore, the extended two-dimensional framework devised by Verhagen et al prescribes that explainability also enables system predictability. It must be noted that this is a theoretical framework, with the aforementioned relations not having been experimentally studied. However, it provides us with a categorization and model that will act as a basis for formulating our hypotheses, determining our independent variable and devising the emergent agent configurations. This current work also acts as a first approach to investigate and validate some of the claims made by Verhagen et al.

When interacting with an incomprehensible system, humans cannot interpret or understand its actions or behaviour, since the agent does not share any kind of information. When the agent becomes transparent, meaning that it discloses "the relevant outward and functional system elements to users, enabling them to access, analyze, and exploit this disclosed information" [2], it is classified as interpretable. This denotes the fact that the agent has reached "the level at which the system's users can assign subjective meanings, draw explanations, and gain knowledge" [2] by making use of this disclosed information. If in addition to disclosure, the agent is also "clarifying disclosed system elements by providing information about causality and establishing relations with other system elements" [2], the discussion shifts towards understandability, which is defined as "the level at which the system's users have knowledge of disclosed and clarified outward and functional system elements, and the relationships and dependencies between them" [2]. An understandable system thus provides explanations for its behaviour. In consequence, subquestions can be formulated regarding the levels of predictability of interpretable and understandable agents and about the difference between them.

Furthermore, since predictability is said to be linked to understandability, we can ask how expertise and experience with the framework within which the agent operates influences its predictability. In order to be able to answer these questions, it must first be established what type of information and knowledge can and should an agent share with its human teammate.

First of all, in the definitions by Verhagen et al presented earlier, functional system elements are mentioned. These refer to aspects such as the agent's world knowledge, intentions, actions, goals, decisions [2]. Secondly, Klein et al argue that an agent must make a number of its system elements clear in order to be perceived as predictable. These are elements like its targets, states, capabilities, intentions and upcoming actions [8]. Finally, in a similar study conducted by Li et al, the following information elements have been identified: the next target, if the agent is requesting assistance, if the agent is providing assistance, other agent's current task, information about task completion, block location, room occupancy and other agent's state [9]. In a similar manner, Harbers et al identified the following two categories of information shared: world knowledge and intentions. Information about the properties of the blocks (such as color, shape, location) and about the agent's own state pertain to the category of world knowledge; its intentions include data about "where the agent is going and which blocks it is going to deliver" [10]. Therefore the following informative elements have been identified relating to the agent operating withing the BW4T context: goals, world knowledge and intentions. Figure 1 details what concrete data is linked to which category.

Since explanations are what transforms an interpretable agent into an understandable one, it was necessary to clearly formulate the explanations the agent present in the ex-

Figure 1: Categorization of information shared by agent

periment would provide to the participants. A list containing various explanations for the actions of the agent has been compiled, with each item being labelled as belonging to one of three categories: linking actions to goals, contrastive or attributive. Whilst it is clear from the naming what kind of explanations are of the first type, a definition will be provided for the other two types. Contrastive explanations are framed with regards to the alternative options, making it clear why this particular action was chosen instead of a different one. Attributive explanations present why that particular action was chosen by virtue of its properties, attributes.

In order to select the best explanations from the compiled list, we have consulted the literature to identify what properties an explanation should have, what criteria it should meet to be deemed as good. Broekens et al claim that "individuals prefer explanations in which intent is communicated by intelligent agents" [11] and Miller argues that "simpler explanations - those that cite fewer causes - and more general explanations - those that explain more events -, are better explanations" [12]. Furthermore, it has been shown that "a good explanation considers both the fact (output) and the foil (alternative output)" [13]. Moreover, in a study conducted by Miller et al, the preference of participants towards simple explanations with less causes and that explained more events became apparent [12]. Thus the criteria of *coherence*, *simplicity*, *generality*, *truth* and *probability* have been applied. The resulting explanations, color coded according to their type, are presented in Figure 2.

3 Block Worlds 4 Teams

To conduct the controlled experiment, the Block Worlds for Teams (BW4T) environment (realized through Matrx ("Human-Agent Teaming Rapid Experimentation software package")) was used. BW4T is a "testbed for team coordination" [10] and consists of the following task: a predetermined sequence of blocks has to be collected and dropped off in the correct place in the specified order. Both agents and humans can act as players, they can communicate with each other through sending messages and through successful collaboration, the team effectiveness can increase. Restrictions are the fact that players can only carry one block at a time and can only see two units around them. Figures 3 and 4 show the so-called god view, where everything is visible, and the agent view, which corresponds to what both computer- and human agents see.

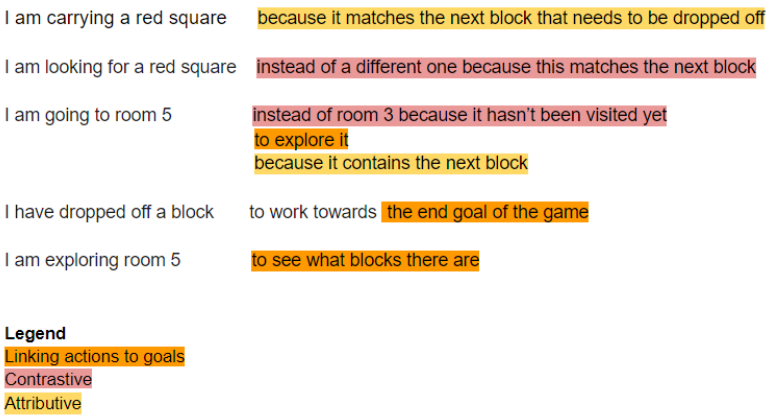

Figure 2: Explanations paired with agent's actions, color-coded according to their type

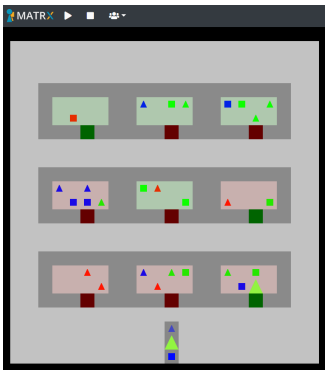

Figure 3: God view in BW4T

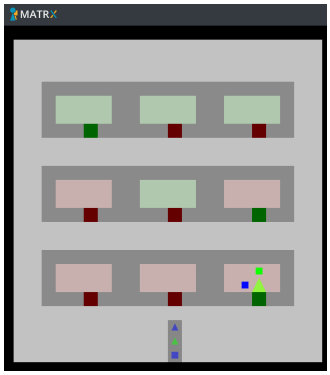

Figure 4: Agent view in BW4T

3.1 The agent

The first version of the agent has been developed within the Collaborative Artificial Intelligence course at the TU Delft. The agent is capable of successfully solving the task on its own, but it can also collaborate with other players. It employs a simple yet effective strategy. Immediately as the task is started, the agent gains knowledge about the blocks that need to be delivered, the so-called Collect blocks. It then starts to explore the yet unvisited rooms in a random order. As soon as it gains knowledge about the location of the Collect block that is the next in the sequence, it will go to the specified location, pick up the block, carry it to the dropoff zone and drop it off. It can gain this knowledge in two ways: it either finds the block by itself while exploring rooms or a teammate has shared the information. If it knows about multiple locations of the same block, it will go to the closest one. In case the agent receives a message about someone else reserving a block, it will start looking for the next block that needs to be dropped off.

According to the categorization of information shared in section 2, the following four agent configurations have been designed: agent sharing only world knowledge, only actions,

world knowledge and actions, and world knowledge, actions and explanations. Furthermore, the first three agents fall into the interpretable category, whilst the last one is understandable. Figure 5 presents the aforementioned configurations.

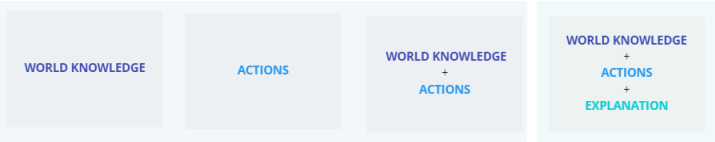

Figure 5: Four agent configurations used in the experiment

According to the configuration, the agent sends messages sharing the corresponding information. For example the first agent will update its teammates about the Collect blocks, about the blocks that it has seen and about the rooms it visited. The third agent will also share this, with the addition of information about what it currently does or will soon do (which room it will go to, which block it is looking for, that it has dropped off a block etc). An example of the messages, as they are visible to the human participants, can be seen in Figure 6.

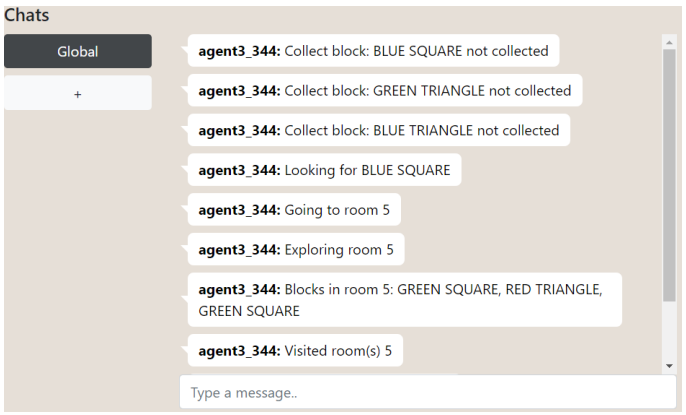

Figure 6: Example of information shared by the agent, as visible to the participant

3.2 Human

In addition to the BW4T environment shown in Figure 4, the human also has access to the chat, where it can see the incoming messages from the agent and where it can also send its own updates. A view of the chat and the corresponding buttons can be seen in Figure 7. Participants can share two kinds of information with the agent: reservation and collection. If participants find a block that they want to drop off, they must first notify the agent of this intention. This is done through sending a *reserving* message. With the use of the drop-down menu, the color, shape and location of the block can be selected. The same procedure also applies when participants have dropped off a block. A *collected* message is sent and thus their teammate is notified of the collection of the block.

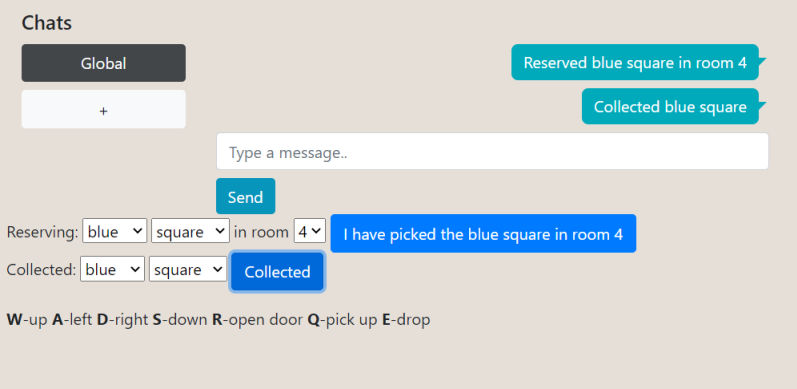

Figure 7: Chat view of participants

4 Experimental Setup and Results

To explore the effect of sharing various types of information on predictability, a controlled experiment was set up, within which participants collaborated with an agent using the BW4T test bed. In order to mitigate the limitations posed by the learning effect, a between-subject design was chosen.

The 27 participants were aged between 20-25, with one participant aged 30 and have various levels of experience with the framework used. The majority, 59.26%, come from a computer science background, but have no previous experience with BW4T, 11.11% on the other hand, have worked with this framework before. The remaining 29.63% have no experience with computer science.

4.1 Variables

The independent variable is the information shared by the agent, giving rise to the four agent configurations detailed in section 3. Regarding the amount of information shared, agents 1 and 2 share a similar amount, with agents 3 and 4 successively increasing this quantity. The dependent variable is the predictability of the agent and team effectiveness. Confounding factors can be the exploration strategy of the agent (for example the fact that initially rooms are chosen at random) and the speed at which messages appear (participants might find it distracting or overwhelming to have to keep up with the flow of updates while concurrently trying to solve the task [9]).

4.2 Measures

Below follows an outline of the measures used with regards to the aforementioned variables.

The percentage of correct predictions a participant achieves will be used as an objective measure of predictability. Part of a questionnaire containing five statements (such as "I feel that the agent was predictable" or "I feel that the agent's actions were consistent") scored on a 5 point Likert scale will serve as a measure of the participant's perception of the agent's predictability. The questions are taken from [14] and have been reformulated to specifically refer to predictability (as opposed to *unpredictability*).

The time taken to complete the task will be recorded and used as the main measure of team effectiveness. Additionally, the number of moves, number of messages sent by both agent and participant and length of messages (expressed in number of words per sentence) will be documented and will be used for gaining further insight into the relationship between predictability and team effectiveness.

The participants who were paired with the agent also sharing explanations will be offered additional questions to gain insight into what role the type of the explanation plays on predictability and to receive understanding of participant's preference regarding the various explanations. The questions are directly taken from the Explanation Satisfaction Scale presented in [15].

4.3 Controlling for the confounding factors

The items shown in Table 1 were asked in order to measure the effect of the outlined confounding factors and to gauge the general attitude of the participants towards the collaboration with the agent. The first and last one are scored on a 5 point Likert scale, while the scoring of the second and third one is the following: 1 - too little, 3 - right amount, 5 - too much.

"I feel like I could work effectively with the agent"
"The amount of updates the agent gave was appropriate"
"The time between messages was appropriate"
"I think that the strategy of the agent was optimal"

Table 1: Scale for measuring effect of confounding factors

4.4 Procedure

Agents are distributed randomly among the participants, who are asked to fill in the consent form. Then participants play one round of BW4T without the agent, to get accustomed to its workings, learn how to navigate and send messages, after which follows the round teamed with the agent. Within this round the task will be paused at various moments and the participants are asked one of the questions accordingly, as outlined in table 2. The

Moment of pause	**Question**
Before the agent picks up a certain block	Which block will the agent pick up next?
Before the agent walks to a room	Which room will the agent go to?
Before the agent drops off a block	What will the agent do with the block?
Before the agent starts looking for the next block	Which block will the agent look for next?

Table 2: Questions the participants were asked, according to when the task was paused

responses are noted, together with the actual action of the agent, observed after resuming the task. After the completion of the task, participants are asked to fill in a questionnaire.

4.5 Pilot experiment

To ensure that the experimental setup is optimal and to mitigate any errors or issues that would render the results unusable, we have undertaken a pilot experiment. Due to the rather restricted scope of this research project, only two participants were selected for it. They were middle aged, with no technical background and had no previous information about the research topic or the experiment.

Valuable insight and feedback has been gained through the running of the pilot. Firstly, it became apparent that the speed of the agent is considerably higher than that of the human participant, with this difference negatively affecting the participant. It lead to a general state of confusion and the whole task has been perceived as overwhelming. As a consequence of the speed of the agent, the messages it sent also appeared at a rate that has been perceived as too fast. It has been reported that it was difficult to keep up with the messages and that having to focus both on reading them, solving the task and additionally providing their own updates was difficult and distracting.

Secondly, it became apparent that the experimenter must know the playthrough of the task particularly well and the moments when the task is paused must be concretized and well defined beforehand. This is necessary to ensure that all participants are having an experience as similar as possible and that they cannot simply read off the correct action of the agent from its messages.

To mitigate these issues in the actual experiment, we have made changes to the behaviour and speed of the agent. Firstly, the tick duration has been increased to 0.2 and the agent's slowdown has been set to 3. Secondly, the message-sending functionality of Matrx has been altered, to ensure that messages are not displayed in batch anymore, but with a time difference of 2 seconds between them. Finally, the planning presented in table 2 has been devised, to ensure a more methodological approach to using the Sagat situation awareness global assessment technique.

4.6 Results

The main purpose of this paper is to find out the relationship between the type and amount of information shared and predictability. In the rest of this section the results obtained will be presented; the agent configurations will be referred to by numbers from 1 to 4, as laid out in table 3.

Number	**Configuration**
1	Sharing world knowledge
2	Sharing actions
3	Sharing world knowledge and actions
4	Sharing world knowledge and actions and explanations

Table 3: Numbering for corresponding agent configuration

According to the objective measure of predictability, agent 4 was the most predictable, with a score of 89.17% (SD = 12.00). This is in line with our hypothesis stating that an understandable agent is more predictable than an interpretable one. The margin by which agent 4 takes the lead is a very small one however, with agent 2 being rated second, with a score of 88.57% (SD = 19.52). An overview of the results of the objective measures, together

Agent	N	Mean	Std Dev
1	7	82.3814	19.3132
2	7	88.5714	19.518
3	7	85	15.5456
4	6	89.1667	12.0069

Table 4: Results of objective predictability measure

Agent	N	Mean	Std Dev
1	5	4.0857	0.4585
2	5	4.1714	0.6655
3	5	3.8571	0.2673
4	5	4.1333	0.4625

Table 5: Results of subjective predictability measure

with the standard deviation can be seen in table 4. A one-way ANOVA test was used to compare the means of the objective measure: there was no statistically significant difference between the the group means ($F(3, 23) = 0.23397$, $p = 1.31369$).

The scores obtained from the subjective measure are summarized in table 5, and alongside the objective measure in figure 8. Similarly to the previous measure, agents 2 and 4 are the most predictable ones, with a score of 4.17 (SD = 0.67) and 4.13 (SD = 0.46) respectively. Agent number 3 has a surprisingly low score: 3.86 (SD = 0.27). The one-way ANOVA test showed that these results are not significant ($F(3, 23) = 0.363$, $p = 0.7802$).We have calculated the correlation coefficient between the objective and subjective measures of predictability to see if there is a significant relationship between them and if one reflects the other: there is a positive correlation of 0.4651 between these two measures.

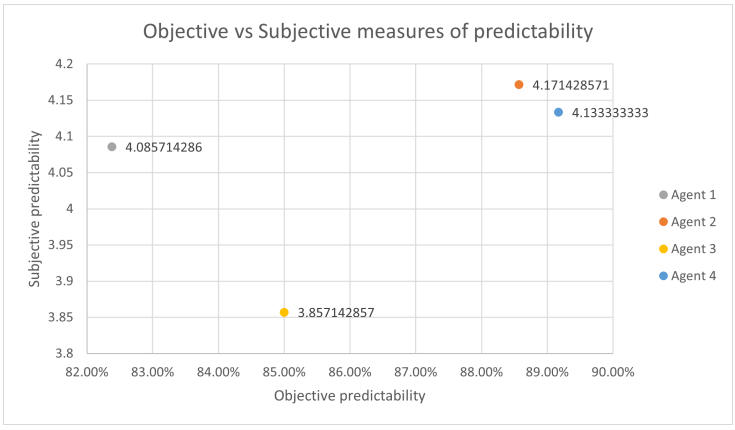

Figure 8: Objective vs subjective predictability

The influence of the participant's experience on (subjective) predictability was also measured. Participants who had previously worked with the BW4T framework gave the highest predictability score: 4.4 out of 5 (SD = 0.0). This was followed by a score of 4.129 (SD = 0.125) granted by people with a background in computer science but no knowledge of BW4T, thus placing the score provided by those with no prior computer science experience, 3.8 (SD = 0.356), in last position. Figure 9 presents these findings below. A one-way ANOVA revealed that there was a statistically significant difference between at least two of these groups ($F(2, 8) = 5.907$, $p = 0.0266$). The post-hoc Tukey's HSD Test found that there is a significant difference between the average predictability scores assigned by participants with no computer science knowledge and those who have previous experience within this

field and with BW4T ($p = 0.023$). The relationship between experience and given score is also supported by a correlation coefficient equal to 0.998.

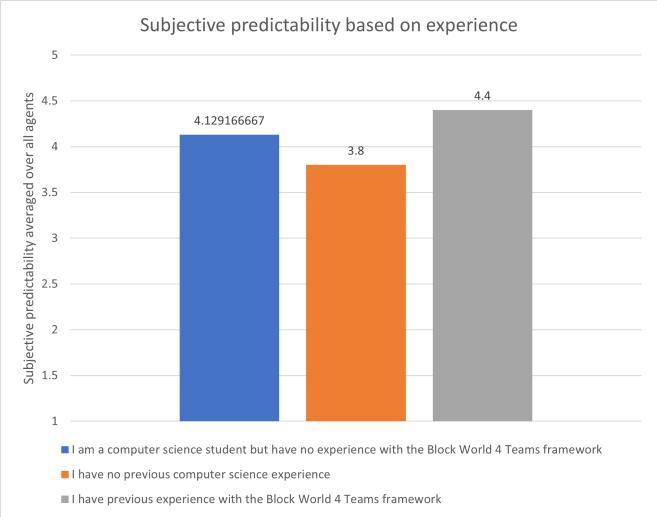

Figure 9: Experience level of participants vs subjective predictability

If we now turn to the measure of team efficiency, we find that the teams containing agent 1 were able to complete the task in the smallest amount of ticks: 733.71 (SD = 245.11), closely followed by agent 3 with a result of 737.71 ticks (SD = 189.88). Participants collaborating with agent 2 amassed an average time of 766 ticks, and finally, the team of agent 4 needed the longest: 919.83 (SD = 354.22). There is a relatively strong positive relationship between efficiency and objective predictability, as demonstrated by a correlation coefficient of 0.7198. The same can not be applied to subjective predictability, since its correlation coefficient with regards to efficiency is only 0.4301.

Predictability seemed to be unaffected by the addition of explanations: agent 4 scored close to agent 2 in both the subjective and objective measure. As shown in figure 10, analyzing the results of the Explanation Satisfaction Scale, we have found that the category with the highest score is that of attributive explanations, closely followed by the contrastive ones and finally, by the category of ones linking the agent's actions to goals. The differences between means are not statistically significant, as shown by a one-way ANOVA test ($F(2, 4) = 3.286$, $p = 0.1432$).

Participants were also asked about the optimality of the agent's strategy, about the amount of messages and time between them and whether they could work effectively with it. The average of their ratings is presented in figure 11.

5 Discussion

This study set out with the aim of determining what type of information shared by an agent leads to the highest predictability. All four agents have been deemed relatively predictable by participants, with scores of above 82% and 3.8. This means that in a simplistic setting with a straightforward and intuitive task, the type of information shared does not bring an added

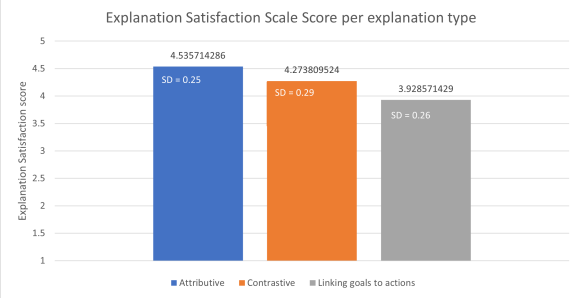

Figure 10: Explanation Satisfaction Scale score per category

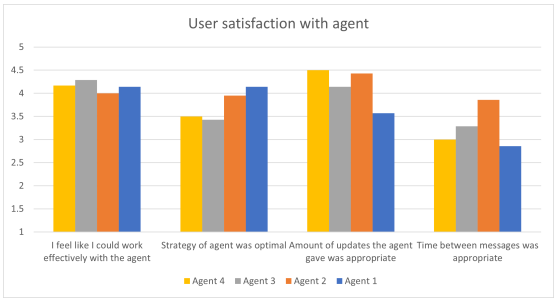

Figure 11: User satisfaction scores

value with regards to predictability. Regarding the categorization of agents as interpretable or understandable, we cannot draw concrete conclusions. The effect of adding explanations (enabling explainability) was not apparent based on our results; no significant differences were observed between the agents pertaining to the first category and the understandable one. Nonetheless, the presence of explanations lead to the highest objective predictability.

Within the current setup, the average score of the appropriateness of the time between messages was 3.25 (SD = 0.44), meaning that participants found it, on average, optimal. Regarding the amount of messages, the tendency of agents sharing more information being scored higher (meaning that participants found the amount too high) is in accordance with the findings of Li et al, who concluded that "too much explanation can hinder a human players ability for decision making" [9]. This idea is further supported by the results stating that teams containing agent 4 took the longest to complete the task. The extra time might have been taken up by the participants reading and processing all of the information and explanations shared by the agent. Moreover, the optimality of the agent's strategy according to the participants tends to diminish as more information is shared and the longer it takes to complete the task. It can be said, that in such a simplistic setup, sharing only world knowledge leads to the highest participant satisfaction with the effectiveness and optimality of the agent: 4.14 out of 5.

It is interesting to take a look at the resulting difference between the agent that is the most predictable and the one that humans are most satisfied with. Whilst the variance in both the predictability and user satisfaction scores is quite low, the latter shows a clear tendency of decrease as the amount of information shared increases. This shows that the amount of information shared has a greater impact on the user experience than only on predictability. Thus it can be said that in a simplistic setting, with a relatively straightforward task, sharing less information will lead to better user experience whilst still achieving a satisfactory level of predictability.

The most prominent finding was the strong influence of previous experience on predictability. This is also in accordance with the work of Klein et al in [8]. Throughout all four agent configurations, the subjective predictability was strongly correlated with participant's experience level and was almost equal across the four cases. This implies that more emphasis should be placed on the type and amount of information shared in the cases when the agent will collaborate with inexperienced users.

5.1 Limitations and Future Work

The scope of this study was limited by several factors. Firstly, the number of participants, 27, is too small to be able to obtain data that can be generalized to the greater population. Secondly, they were not diverse enough: only 29.63% came from a non-technical background and the majority were aged 20-25. The two main external factors influencing the choice of participants were the COVID-19 pandemic and the timeframe of the research project (10 weeks). Nevertheless, the results obtained still provided us with useful knowledge about the different factors at play with regards to predictability and team effectiveness in human-agent teams.

The world and task used in BW4T are suspected to have been too simple and intuitive. Participants were able to correctly predict the agent's next action even when no information has been shared yet. The reason for having a more simplistic world was the expectation that participants will find it overwhelming to navigate a more complex world whilst also processing the information shared by the agent and they themselves sending messages of their own. Still, these results gave valuable insight into how the simplicity of the environment and task influence predictability, and how, in such cases, it is more worthy to focus on devising agents leading to increased user satisfaction.

We have considered the strategy of the agent a potential confounding factor. The average score of 3.75, with a standard deviation of 0.35 shows that participants scored the strategy very similarly across the four cases, meaning that this factor did not significantly influence the results of any of the configurations compared to the others.

The presentation and amount of messages can be optimized in the future. One participant suggested that alongside visually displaying the information shared by the agent, it could also be presented in an audio format. This would remove the need of the human to take its eyes and focus off the task and read the messages. Another suggestion was the replacement of the dropdown menus used to select the reserved or collected blocks with more visual cues, such as icons representing the shapes and colors.

In future investigations it might be interesting to have a more complex world set up in BW4T. This could involve more rooms, more blocks to be collected, the necessity of collaborations between human and agent. More complexity would also mean a longer duration, thus more opportunities for the agent to share information. This would result in more data obtained. However, as previously mentioned, it must be ensured that there is enough time between the messages as to not overwhelm the participants.

To establish the effect of explanations on predictability and to gain valuable and significant knowledge about the influence of different types of explanations, further research is required. A separate study can be set up to this end. Possible questions to pose could be whether a specific type of explanation leads to higher predictability, if a combination of categories is preferred by people or they could regard the relationship between the increase in predictability due to the presence of explanations and the mental effort required from the humans to process the explanations.

5.2 Responsible Research

In this subsection we will discuss the ethical implications and considerations relating to this research and experiment. Firstly, the Human Research Ethics Committee (HREC) of the TU Delft has approved the study and has deemed it risk free. This means that there are no known risks associated to taking part in the experiment. This is clear, since for example no personal or sensitive data of the participant is recorded or stored, none of the participants

are part of vulnerable groups, nor are they in a subordinate position with the experimenter. Moreover, all the participants have been informed about the scope and contents of the study and have signed the informed consent form. Due to the nature of the experiment, it was not necessary to deceive participants.

The data obtained through conducting the experiment is used only for the purposes of answering the posed research question and to gain insight into the topic of the research. There is a possibility, however, for it to be used in future research relating to this domain. All measures have been taken to ensure a correct handling of the data and mitigation of biases. The data has been analyzed in an objective manner, using statistical tests. All data points have been included in the analyses and participants got assigned an agent in a randomized manner. There is no conflict of interests present in the context of this research.

An important note and consideration when interpreting the data nonetheless is the fact that it is not representative, nor balanced. First of all, the participants are not representative of the general population. The majority of them have a background in computer science, are students in their twenties and know the experimenter personally. Secondly, the findings are not compared to a baseline consisting of data about information sharing and predictability in human-human teams. This means that nothing concrete can be stated about the generalizability of the results or whether they relate to findings from the social sciences about predictability or role of information sharing in human teams. The findings can give insight into these topics in the context of teams composed of agents and students in their twenties with some technical background. In order to obtain better data, the experiment must be reproduced with a much more diverse, representative participant group.

The aforementioned issues further highlight the importance of reproducibility in academia. To ensure the reproducibility of our study, we took several measures. Firstly, a detailed description of the experimental setup, variables and measures is provided. A list of potential confounding factors is also identified, together with the measures we took to mitigate their effect. Secondly, the code behind the agent and experimental environment is available upon request to anyone willing to work with it. We believe that it is of utmost significance for researchers to actively take steps to ensure the reproducibility of their studies and experiments.

6 Conclusion

The purpose of this research was to investigate the influence of information sharing on predictability in human-agent teams. As presented in the previous sections, sharing information and explanations is what makes an agent understandable. Understandable agents are deemed to be more predictable, which in turn is supposedly increasing trust between the human and agent. Trust in such relationships is of utmost importance, since it also leads to an increased team efficiency.

We have conducted a controlled experiment with four agent configurations: sharing world knowledge, actions, world knowledge and actions, or world knowledge, actions and explanations. Whilst none of the results obtained are of statistical significance, they do offer valuable insight into how information shared ties into predictability. Our main findings show that there is a strong positive correlation between previous experience with a given framework or agent and perceived predictability. However, the difference in predictability between the four different agents is minimal, so no clear conclusion can be drawn about what information should an agent share to achieve maximal predictability. Sharing more information has lead to a longer time to complete the task and participants found that the

agent sharing everything provided too much information. Moreover, the longer it took to complete the task, the lower the rating of the agent's strategy. Thus in this simple world setup in BW4T, participants were most satisfied with working with the agent sharing only world knowledge.

The most important limitation lies in the limited number and diversity of participants and in the simplicity of the BW4T world used. Therefore the experiment should be repeated in a manner that mitigates these limitations and with improvements suggested in the previous section in mind. Future research might explore more in depth the effect of the different types of explanations on predictability or analyze explicitly the relationship between predictability and trust.

7 Acknowledgements

The author would like to thank Dr. Myrthe Tielman, Ruben Verhagen and Carolina Jorge for their guidance, feedback and support, and for providing the opportunity for gaining valuable insight into the world of research and academia. Yuxin Jiang, Stephen van der Kruk and Ziad Nawar for their support, collaboration and suggestions for improvement. Wout Haakman, Jacob Roeters van Lennep and Kristof Vass for their contribution in developing the first version of the agent in the Collaborative AI course. Marton Soos for his continuous support and helpful suggestions. And all the participants for taking time to partake in the experiment.

References

[1] J. Bradshaw, V. Dignum, C. Jonker, and M. Sierhuis, "Human-agent-robot teamwork," pp. 487–487, 2012.

[2] V. Ruben S, M. Neerincx, and M. Tielman, "A two-dimensional explanation framework to classify ai as incomprehensible, interpretable or understandable," 2021.

[3] M. Johnson and A. Vera, "No ai is an island: the case for teaming intelligence," *AI Magazine*, vol. 40, no. 1, pp. 16–28, 2019.

[4] M. Johnson, J. Bradshaw, P. Feltovich, C. Jonker, M. van Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 43–69, 2014.

[5] K. Stubbs, P. J. Hinds, and D. Wettergreen, "Autonomy and common ground in human-robot interaction: A field study," *IEEE Intelligent Systems*, vol. 22, 3 2007.

[6] M. Johnson and J. T. Bradshaw, "The role of interdependence in trust," 2021.

[7] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. FrÃ€mling, "Explainable agents and robots: Results from a systematic literature review," pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[8] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, "Ten challenges for making automation a "team player" in joint human-agent activity," *IEEE Intelligent Systems*, vol. 19, pp. 91–95, 2004.

[9] S. Li, W. Sun, and T. Miller, "Communication in human-agent teams for tasks with joint action," in *Coordination, Organizations, Institutions, and Norms in Agent Systems XI* (V. Dignum, P. Noriega, M. Sensoy, and J. S. Sichman, eds.), (Cham), pp. 224–241, Springer International Publishing, 2016.

[10] M. Harbers, B. V. Riemsdijk, and C. Jonker, "Measuring sharedness of mental models and its relation to team performance."

[11] J. Broekens, M. Harbers, K. Hindriks, K. van den Bosch, C. Jonker, and J.-J. Meyer, "Do you get it? user-evaluated explainable bdi agents," in *Multiagent System Technologies* (J. Dix and C. Witteveen, eds.), (Berlin, Heidelberg), pp. 28–39, Springer Berlin Heidelberg, 2010.

[12] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[13] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Engineering Psychology and Cognitive Ergonomics* (D. Harris, ed.), (Cham), pp. 204–214, Springer International Publishing, 2018.

[14] B. R. Schadenberg, "Understanding and facilitating predictability for engagement in learning," 4 2021.

[15] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects institute for human and machine cognition," 2018.

A Questionnaire used

Please pick the option that most accurately describes your current situation:

- ◯ I am a computer science student but have no experience with the Block World 4 Teams framework
- ◯ I have previous experience with the Block World 4 Teams framework
- ◯ I have no previous computer science experience

I feel that the agent was predictable

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

I feel that the agent's actions were regular

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

I feel that the agent's actions were consistent

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

I feel that the agent was not erratic

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

I feel that the agent did not act randomly

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I feel like I could work effectively with the agent

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The amount of updates the agent gave was appropriate

	1	2	3	4	5	
The agent gave too few updates	○	○	○	○	○	The agent gave too many updates

The time between messages was appropriate

	1	2	3	4	5	
There was too little time	○	○	○	○	○	There was too much time

The strategy of the agent was optimal

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Explanations

For each explanation shown, please indicate to what extent you agree with the statements presented (1 = Strongly disagree, 5 = Strongly agree)

"I am carrying a read square because it matches the next block that needs to be dropped off"

	1	2	3	4	5
From the explanation, I understand how the agent works	○	○	○	○	○
This explanation of the agent's action is satisfying with regards to predictability	○	○	○	○	○
This explanation of how the agent works has sufficient detail	○	○	○	○	○
This explanation of how the agent works seems complete	○	○	○	○	○
This explanation of the agent's action is useful to my goals	○	○	○	○	○
This explanation of the agent shows me how accurate the agent is	○	○	○	○	○
This explanation lets me judge when I should trust and not trust the agent	○	○	○	○	○

"I am looking for a red square instead of a different one because this matches the next block"

	1	2	3	4	5
From the explanation, I understand how the agent works	○	○	○	○	○
This explanation of the agent's action is satisfying with regards to predictability	○	○	○	○	○
This explanation of how the agent works has sufficient detail	○	○	○	○	○
This explanation of how the agent works seems complete	○	○	○	○	○
This explanation of the agent's action is useful to my goals	○	○	○	○	○
This explanation of the agent shows me how accurate the agent is	○	○	○	○	○
This explanation lets me judge when I should trust and not trust the agent	○	○	○	○	○

"I am going to room 5 to explore it"

	1	2	3	4	5
From the explanation, I understand how the agent works	◯	◯	◯	◯	◯
This explanation of the agent's action is satisfying with regards to predictability	◯	◯	◯	◯	◯
This explanation of how the agent works has sufficient detail	◯	◯	◯	◯	◯
This explanation of how the agent works seems complete	◯	◯	◯	◯	◯
This explanation of the agent's action is useful to my goals	◯	◯	◯	◯	◯
This explanation of the agent shows me how accurate the agent is	◯	◯	◯	◯	◯
This explanation lets me judge when I should trust and not trust the agent	◯	◯	◯	◯	◯

"I am going to room 5 instead of room 3 because it hasn't been visited yet"

	1	2	3	4	5
From the explanation, I understand how the agent works	○	○	○	○	○
This explanation of the agent's action is satisfying with regards to predictability	○	○	○	○	○
This explanation of how the agent works has sufficient detail	○	○	○	○	○
This explanation of how the agent works seems complete	○	○	○	○	○
This explanation of the agent's action is useful to my goals	○	○	○	○	○
This explanation of the agent shows me how accurate the agent is	○	○	○	○	○
This explanation lets me judge when I should trust and not trust the agent	○	○	○	○	○

"I am going to room 5 because it contains the next block"

	1	2	3	4	5
From the explanation, I understand how the agent works	◯	◯	◯	◯	◯
This explanation of the agent's action is satisfying with regards to predictability	◯	◯	◯	◯	◯
This explanation of how the agent works has sufficient detail	◯	◯	◯	◯	◯
This explanation of how the agent works seems complete	◯	◯	◯	◯	◯
This explanation of the agent's action is useful to my goals	◯	◯	◯	◯	◯
This explanation of the agent shows me how accurate the agent is	◯	◯	◯	◯	◯
This explanation lets me judge when I should trust and not trust the agent	◯	◯	◯	◯	◯

"I have dropped off a block to work towards the end goal of the game"

	1	2	3	4	5
From the explanation, I understand how the agent works	○	○	○	○	○
This explanation of the agent's action is satisfying with regards to predictability	○	○	○	○	○
This explanation of how the agent works has sufficient detail	○	○	○	○	○
This explanation of how the agent works seems complete	○	○	○	○	○
This explanation of the agent's action is useful to my goals	○	○	○	○	○
This explanation of the agent shows me how accurate the agent is	○	○	○	○	○
This explanation lets me judge when I should trust and not trust the agent	○	○	○	○	○

"I am exploring room 5 to see what blocks there are"

	1	2	3	4	5
From the explanation, I understand how the agent works	○	○	○	○	○
This explanation of the agent's action is satisfying with regards to predictability	○	○	○	○	○
This explanation of how the agent works has sufficient detail	○	○	○	○	○
This explanation of how the agent works seems complete	○	○	○	○	○
This explanation of the agent's action is useful to my goals	○	○	○	○	○
This explanation of the agent shows me how accurate the agent is	○	○	○	○	○
This explanation lets me judge when I should trust and not trust the agent	○	○	○	○	○