Directability's influence on human-agent trust

Yuxin Jiang[1] , **Ruben Verhagen**[1] , **Carolina Jorge**[1] , **Myrthe Tielman**
[1]TU Delft

Abstract

Mutual human-agent trust is of great importance for humans and agents to complete a task collaboratively. This paper aims at studying one of the factors influencing this mutual trust, the directability of humans to agents. Previous studies either take directability as a general concept without looking into its different representations or fail to bridge the gap between directability and trust. This paper starts by analyzing directability's different representations including commands, suggestions, and warnings, then investigates their influences on trust respectively. The experiment is set up in Block World For Teams(BW4T). Afterward, the trust is measured both by calculating the risk-taking behaviours by humans and using questionnaires. The result from the experiment suggests that collaborating with directability improves trust from humans to agents. Among different directability representations, commands and suggestions are the best ways to boost trust. However, the confounding factors such as familiarity with the experiment also make a difference to the final result, those factors can be further investigated in the future.

1 Introduction

Mutual human-agent trust has been a popular research topic in the past few years, previous researchers on this topic have realized that autonomy is hard to achieve when excluding human as potential teammates [1]. Besides, AI can only reach its full potential when woven into human work practice [2]. As a consequence, for humans and agents to work efficiently together, mutual trust is essential. Trust between humans and agents is defined by Lewis as the "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [3 p136]. Studies show that proper trust towards agents can make it more reliable [3] while inappropriate levels of trust will negatively impact team performance, for example, a high level of trust can potentially lead to over-reliance and misuse. A low level of trust, however, may bring to the disuse of the system [4].

Many studies have been conducted in this field to study the factors which contribute to better human-agent trust. Those factors can be classified as human-related, agent-related, and environmental-related [4]. For instance, the level of automation for an agent is an agent-related factor, while attention capacity is a human-related factor. Moreover, the interdependence relationship between humans and agents including observability, predictability, and directability also plays an important role in affecting mutual human-agent trust [5], but some of them haven't been studied thoroughly before. This research focuses on one interdependence relationship - directability, to study its impact on human-agent trust.

The research question is formed as below: "Does direactbility of an agent improve trust in that agent?" According to Johnson and Vera [2], directability is defined as "one's ability to direct or influence the behaviour of others and complementarily to be directed and influenced by others", in this research, the emphasis is on a human's ability to direct and influence an agent's behaviour. To be more specific, the research investigates different ways in which a human can direct agents, including sending commands, suggestions and even warnings. Our hypothesis is, directability of a human to an agent improves trust in that agent. Moreover, the more direct or more mandatory the order is, the more trust will be given to the agent. In other words, we assume that commands would give the most trust from the human to the agent

Previous studies on this topic mainly take the directability as a whole concept without differentiating its various representations. However, we are curious about whether different representations of directability will have different effects on trust, and to test the hypothesis, the main research question is broken down into four sub-questions. Firstly, since one of the two keywords in the research question is directability, the first two sub-questions are formed as: "what is directability?", "what are the different representations of directability?" Resolving those two questions can help to obtain a more comprehensive understanding of directablity. Then the third sub-question focuses more on the second keyword trust. Since trust is a subjective feeling, it requires various ways to measure it concretely, thus the sub-question is formed as: "what are the different ways to measure human-agent trust?" Finally, the relationship between those

two keywords is required to be studied as well, so we propose the last sub-question as: "How do different representations of directability contribute to the human-agent trust?"

The report is structured as follows, section 2 presents the literature review and our main contribution to this research. Then, the detailed experimental setup follows in section 3. The results of the experiment can be found in section 4. Section 5 contains the ethical aspect of our research and discusses the reproducibility of our methods. Based on that, a more comprehensive discussion can be found in section 6, followed by the limitation and future work in section 7. The whole report is concluded in section 8.

2 The contribution on directability and trust measurement

2.1 Gap analysis on directability

In the first place, we will present our literature review and gap analysis on directability. Previous studies show that in different contexts, directability could have different representations. For example, Myers and Morley [6] presented a framework that involves adjustable autonomy and strategy preference as representations of human's directability to agents. Johnson and Bradshaw [7] proposed several specific representations of directability including commands, suggestions, warnings, progress appraisals, helpful adjuncts. However, when it comes to their influence to trust, they took directabilty as a whole concept, stating that directability allows the trustor to take the initial steps of partial trust [8]. All in all, although some researchers have investigated different representations of directability, few of them further analyzed their influence on human-agent trust. At the same time, the papers which did look into the relationship between directability and trust, however, tend to ignore directability's multiple representations.

Paper	Representations of directability	Directbility's influence on trust	Different representations of directability's influence on trust
Myers et al., 2001	✓		
Myers et al., 2003	✓		
Johnson et al., 2021	✓	✓	
Lewis et al, 2018		✓	
This paper	✓	✓	✓

Figure 1: Gap analysis of papers relevant to directability

The main contribution of this paper is bridging the gap between various forms of directability and human-agent trust. To be more specific, The paper selects three representations of directability, commands, suggestions, warnings since they are relatively easier to implement in BW4T, controlled experiments are used to investigate their influence on human-agent trust respectively.

2.2 Gap analysis on measurement of trust

When it comes to trust measurement, previous studies have presented different ways of measuring trust, as shown in Figure 3, they can generally be classified into subjective ways and objective ways. Jian et al. 1998 [9] investigated the difference between human-human trust and human-machine trust by looking into a set of words people give related to trust, the result is that people do not perceive the concepts of trust differently across different relationships. Based on that, they developed their trust scale by starting with a word elicitation task. They extracted a 12-factor structure used to develop a 12-item scale based on the examination of clusters of words [3]. Hoffman [8] summarized several scales used in trust measurement and developed a Hoffman scale which is specifically suitable for the Explainable AI system (XAI). The above papers share one feature in common, they did not treat the trust as one dimension factor, instead, factors including predictability, reliability are covered in the measurement as well. However, as stated by Lewis et al. [3], self-report through psychometric instruments is the most direct measurement while questionnaires still suffer from many weaknesses such as it can only be conducted after the experiment. Recognizing that, Schaefer [4, 10] developed a Trust Perception Scale-HRI, which includes 40 psychometrically-developed items.

Compared to the subjective measurement of trust, the objective measurement was mentioned less in previous papers, the main issue with objective measurement is how to represent mathematically the concept of trust [11]. According to Freedy, a rational decision model is required to obtain a trust score, in other words, we need to assume that participants are rational. Based on that, Schaefer [10] obtained his objective measurement by measuring the percentage of time a human attends to a robot during a human-robot team navigation task.

Measurement of trust	Subjective	Objective
Jian et al., 1998	Empirical Derived 12-item scale	
Hoffman et al., 2018	Hofffman scale	
Schaefer, 2013	Psychometric instruments	time attending to robot
Freedy et al., 2007		rational decision model
This paper	Hoffman scale	risk-taking baheviours (human go to the room with tiger)

Figure 2: Gap analysis of papers relevant to measurement of trust

Although combining subjective measurement and objective measurement of trust is not a first-time breakthrough. It is still worth mentioning that this research innovates by measuring the risk-taking behavior of participants in BW4T. When revisiting the definition of trust, vulnerability and uncertainty are two important elements. Perceived risk and risk-taking in a relationship are also important components in the ABI trust model proposed by Mayer and Davis [12]. As a result, a risky environment is deliberately created in BW4T to measure the trust in a relatively objective way. When combined with the subjective measurement using questionnaires afterward, we obtain a more comprehensive insight into trust.

2.3 Bridging the literature review and the experiment

Until now, the answers to the first three sub-questions have already been covered through a literature review. They are going to be briefly revisited here, which helps us enter the next section. Firstly, the directability in this research is defined as a human's ability to direct or influence the behaviour of the agent. Among its various representations, commands, warnings, and suggestions are selected to further investigate their influences on trust respectively. The measurement of trust is completed through calculating the risk-taking behaviours by human, along with the subjective measurement making use of the modified Hoffman scale [8]. With this knowledge in mind, we are prepared to investigate the fourth sub-question in the next section.

3 Methdology and Experimental Setup

3.1 Background

Block World for Teams

The first three sub-questions are approached mainly through literature review. To answer the fourth sub-question, an experiment was set up in the Blocks World for Teams(BW4T), a joint activity testbed [13]. It is a collaborative game involving a human and an agent, their task is to grab targeted blocks and put them in the drop-zone at the bottom. As shown in Figure 3, the two green triangles in the top left corner represent a human and an agent. There are a total of nine rooms, each of them contains different blocks, the blocks which match the shape and color of the block in the drop-zone are targeted blocks while others are non-targeted blocks. Players can navigate their avatars around the world and perform pick-up and drop-off actions [13]. In the meantime, they can also send messages to the agent through a pre-defined API. When all the targeted blocks are successfully collected, the game is over and the total time taken is recorded. The goal for the team is to collect all targeted blocks as soon as possible.

Implementation of agents

Human and agent: The experiment setup (Figure 5) is based on the basic setup explained in BW4T. the color is used to differentiate the human and the agent, the blue triangle represents the human, its sense range is 5, which means the human can detect blocks within five cells distance. However, it can see all the other agents in the world. Its speed is 1, which means that the human makes one move per tick. The most important feature of the human is sending messages, as in Figure 4, there are a total of three types of messages the human can send, the first is the command, players can go to the chatbox, find agent 1, and type "c0 - c8", which corresponds to giving commands to the agent to go to room 0 to room 8 (room number starts from 0). Similar to the command, when players type "w0-w8" it gives a warning to the agent not to go to certain rooms. And finally, "s0-s8" represents suggesting the agent go to those rooms. On the agent side, for it to be directable, the agent must have a human-usable interface for providing directions and the algorithms must support ingestion and incorporation of directions [7]. In the experiment implementation, all types of

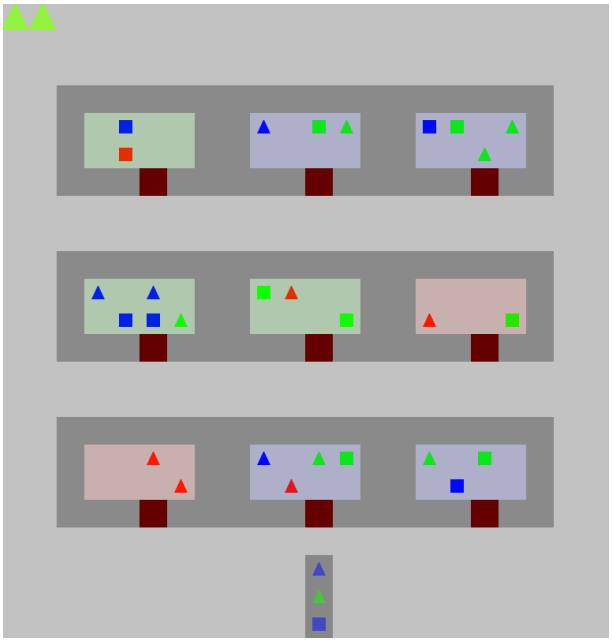


Figure 3: Block World for Teams

messages are automatically stored in the list of the received messages of the agent, each time the agent has to select a targeted room, the list is traversed.

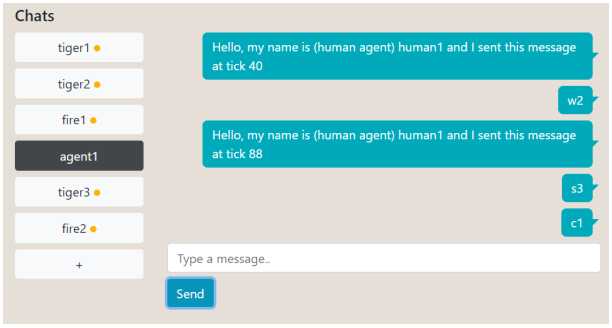


Figure 4: Messages human can send to agent

When looking at the green agent, its speed is 10 times slower than the human and its sense range is one, after the start of the game, it first goes to the drop zone and figures out the next targeted block, it is worth noticing that the agent can have only one target at a time. In other words, it doesn't know its next target block until the previous one is collected. Each time when the agent has to choose a room to head to, it will traverse the list of the received messages from the latest to the oldest. If the next message in the list is a command, it will strictly follow that command, if it is a suggestion, the agent has some probability to follow that suggestion (the probability is not known to the player). In the meantime, a warning list is created from warning messages received, if neither command nor suggestion is found in the list of the received messages, the agent will randomly pick a room

that is not in the warning list as the next target. However, there is also a very small chance that the agent will ignore the warnings and pick a room in the warning list. Besides, we have also implemented the strategy for the conflicting situation, if both warning and command are given for the same room, the warning will be ignored, and the command will take effect. if multiple commands are given for the different rooms, the agent will follow the latest command.

Tiger and Fire: In addition, two extra agents are added to the game, As in Figure 5, the orange circle represents tiger, tiger moves randomly in the room, but when the door is open, it is possible for the tiger to go out of the room it comes from. The moving speed of the tiger is 20 times slower than the human. The most important feature of the tiger is that it can bite and eat humans, when a human comes to the cell around the tiger for the first time, the tiger will bite the human, and the second time, the human will be eaten, and the game is over. However, the tiger can't hurt the agent. The rule for fire is similar, the only difference is that fire can burn both the human and the agent, and the first time the human or the agent is burnt, the game is over.

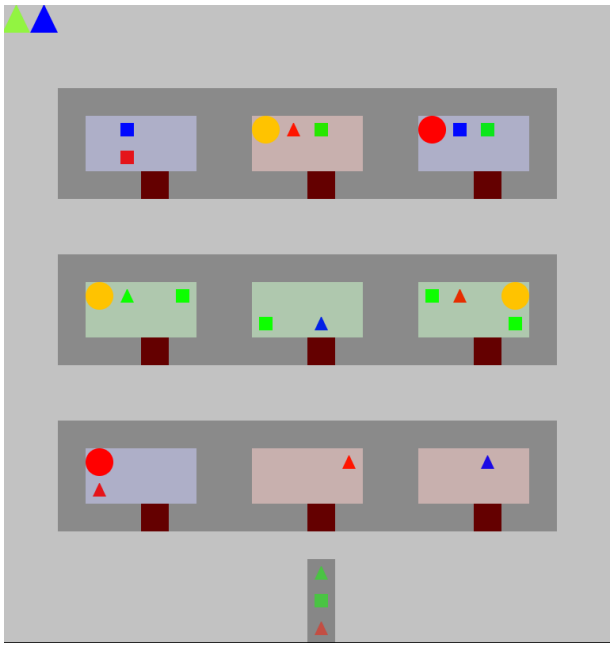


Figure 5: Experiment setup

3.2 Pilot test

In scientific research, pilot testing is a small preliminary study used to test a proposed research study before a full-scale performance [14]. In our case, before the formal experiment, three participants (2 males, 1 female) were invited to join the pilot test. They played the game for four rounds in a fixed order, the first round without directability, the second round with commands, the third round with warnings, and the final round with suggestions. According to the result (Appendix E), suggestions achieved the highest trust score.

However, reflecting on the whole process, it is suspected that the fixed order gives participants more familiarity with the experiment when it comes to the final round. In that sense. more familiarity leads to more trust, which explains why suggestions surpassed the others in trust measurement. To avoid the noise caused by familiarity and make the experiment result more reliable, we decided to shuffle the order of directability representations in the formal experiment, for example, the first player starts with commands, the second starts with suggestions, the third with warnings, etc.

3.3 Participants

During COVID-19 times, it is hard to recruit participants, most of the experiments were completed online, which also brings extra inconvenience to the research. In total, 12 participants (8 males, 4 females) were recruited for this experiment. They have different backgrounds, 10 of them are students between 20 and 25 years old. 2 of them are between 50 and 60 years old. Participants all volunteered to join this experiment, it took them around 40 minutes on average to complete the whole process.

3.4 Procedure

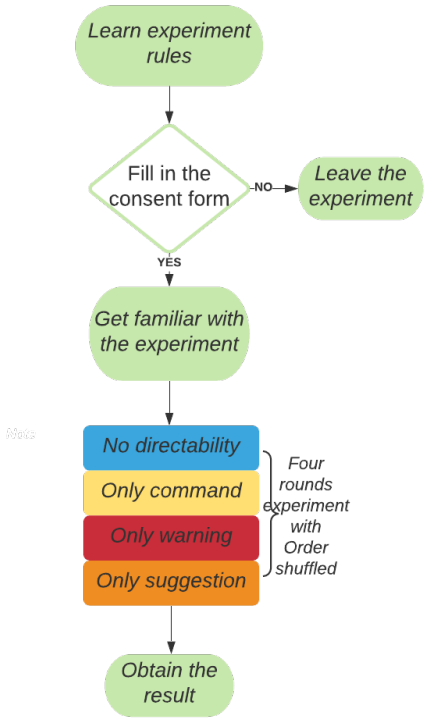


Figure 6: Procedure of experiment

As in Figure 6, the experiments were conducted in the following procedure, first, researchers inform participants about the rules and procedures of the experiment, an information

sheet (Appendix D) and a consent form (Appendix B) are used in the period to ensure the participants are aware of the risks involved and rights they have. With that knowledge in mind, each participant spends some time getting used to the system, this process usually involves two or three rounds of playing to make sure they have a considerable amount of familiarity with the experiment.

After confirming with participants they are ready to begin, all of them are required to play the game for four rounds, and for each round, he or she is either not allowed to send anything or allowed to send only one type of message to the agent. The order for these four rounds is not fixed, the shuffle of the order avoids the potential noise caused by familiarity with the game. After each round, participants are asked to fill in a questionnaire that measures the level of trust they give to the agent.

For each participant, we collect four questionnaires, which means that there are 48 questionnaires in total. After all the participants finished their games, the data from 48 questionnaires were analyzed and compared.

3.5 Measures

Risk taking behaviour

As mentioned previously in gap analysis, perceived risk is an important element in the ABI trust model [12]. We use risk-taking behaviour as the first measurement of trust. it is possible for a human to complete the game by itself without the help of the agent. However, if a human decides to take the risk and relies on the agent to pick up a block instead of fetching it by himself, it represents the human's trust in that agent. In view of that, it appears natural to take the number of commands the human sends to the agent as a measurement of trust. But this measurement is problematic. The reason behind is simple, since the command, as one of the directability representation, is an independent variable, so it can't be taken as a dependent variable at the same time.

Alternatively, the number of times humans entering the room with a tiger is taken as the factor to measure trust. We assume that with a certain level of familiarity with the game and rules, participants are rational in the game. As a result, they are aware that entering the room with a tiger is a risky action, the best strategy human can take to finish the game as quickly as possible is to send messages to the agent and let it do the job for him. However, if the human decides not to do that, if he enters the room with the tiger by himself, it represents not enough trust given to the agent. Overall, the more times the human enters the room with a tiger, the less trust it gives to the agent.

Compared to questionnaires, calculating the risk-taking behaviours doesn't ask participants' opinions directly, it seems that it is an objective method to measure trust. However, the decision a human makes on whether to enter the room with the tiger is still a personal judgment. Although it is assumed that participants are rational, we can not safely assert that the behaviour is not influenced by personal feelings or opinions. As a result, this measurement helps to obtain a more comprehensive view of trust, but can not act as a hundred percent objective measurement and explain everything.

Questionnaires

The subjective measurement of trust is completed through the questionnaire filled in by the participants after the game. The questionnaire is developed from Hoffman's scale (Appendix A). The Hoffman scale is composed of eight questions in total, for each question, there are five scales available from strongly disagree to strongly agree. The factors that are covered including predictability, reliability, and efficiency. We chose the Hoffman scale since it is designed for the human-agent context. Besides, it combines several scales, which are all tested to be highly reliable [8]. To get a trust score, apart from the sixth question: I am wary of the agent, whose score needs to be reversed (4 becomes 2, 5 becomes 1, 3 unchanged), the answers to other questions are added up and divided by the total number of questions, eight. The average score for those eight questions is taken as the subjective measurement of trust.

3.6 Controlled experiments

In conclusion, the experiments were conducted in a manner of controlled experiments. The important elements of controlled experiments are listed here:

Independent:
The type of messages humans can send to the agent, there a total of four, send nothing, send commands, send suggestions, send warnings.

Dependent:
Average score of Hoffman scale
The number of times human goes to the room with a tiger.

Potential Confounding factors:
Familiarity with the game, Demographic information, Team performance

4 Results

As mentioned previously in the measures section, there are 8 values (8 columns) obtained for each participant, the questionnaire score, and the count of risk-taking behaviours for four representations of directability (Appendix C). For each measurement, we calculated their sum, average, and standard deviation(SD), Besides, we also counted the number of winning rounds (the figures marked in green) in total for these four types of representations.

4.1 Questionnaires measurement

In subjective measurement, a box plot was drawn to show the results (Figure 7), commands give the highest trust score with a 3.10 average, suggestions come next with a small gap at 2.94 while warnings give 2.69 trust scores. It is illustrated

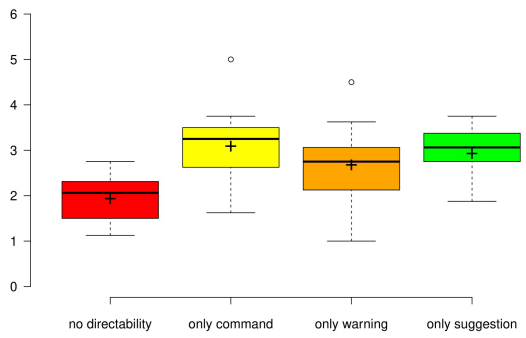


Figure 7: Box plot for subjective measurement

Box plot statistics

	no directability	only command	only warning	only suggestion
Upper whisker	2.75	3.75	3.62	3.75
3rd quartile	2.31	3.50	3.06	3.38
Median	2.06	3.25	2.75	3.06
1st quartile	1.50	2.62	2.12	2.75
Lower whisker	1.12	1.62	1.00	1.88
Nr. of data points	12.00	12.00	12.00	12.00
Mean	1.95	3.10	2.69	2.94

Figure 8: Box plot statistics

from the table that without directability, the trust score of 1.95 is lower than the others.

Besides, from the box plot, the datasets for commands and warnings are more dispersed compared to suggestions and without directability. It is also verified by the SD (standard deviation) calculated, the SD for the no directability is 0.51 while the suggestion is 0.56. The figures are lower than the SD for the command (0.86) and the warning (0.85).

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	9.381184896	3	3.127061632	5.592467143	0.002443678091	2.816465827
Within Groups	24.60286458	44	0.5591560133			
Total	33.98404948	47				

Figure 9: ANOVA analysis of subjective measures

Apart from the box plot, an ANOVA analysis was also conducted on the subjective measurement's result. ANOVA is a statistical technique that is used to compare the means of more than two populations [15]. There are two important hypotheses involved in ANOVA, null hypothesis(H0) and alternative hypothesis(H1). The null hypothesis suggests there is no significant difference between sample means, it holds when p-value is larger than 0.05. The alternative hypothesis suggests the opposite, and it holds when p is smaller than 0.05. According to Figure 9, the p-value we get is 0.002, telling that a significant difference exists in the dataset.

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inferfence
A vs B	5.3564	0.0024994	** p<0.01
A vs C	3.4262	0.0875015	insignificant
A vs D	4.5843	0.0117308	* p<0.05
B vs C	1.9302	0.5249790	insignificant
B vs D	0.7721	0.8999947	insignificant
C vs D	1.1581	0.8270011	insignificant

Figure 10: Turkey HSD analysis of subjective measures

However, even if we are aware that there is a significant difference in the dataset, we do not know the difference is in which two representations. So a further test is required to compare the difference between every two representations and identify the ones that are greater than the standard error. So we applied Turkey HSD to our sample. In Figure 10, A represents no directability, B represents only commands, C represents only warnings, D represents only suggestions. As we can see, there is a significant difference between no directability and commands, no directability and suggestions. Except for those two, the difference is insignificant.

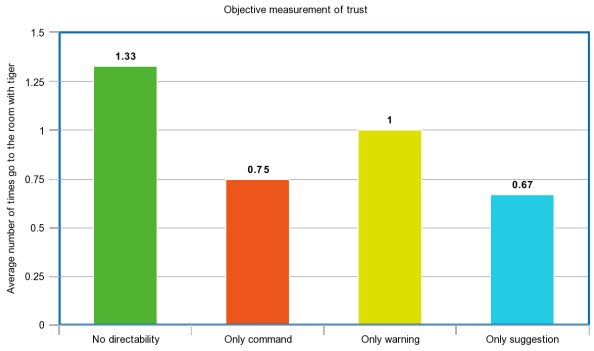


Figure 11: Objective measurement of trust

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3.229166667	3	1.076388889	2.693522907	0.05753716323	2.816465827
Within Groups	17.58333333	44	0.3996212121			
Total	20.8125	47				

Figure 12: ANOVA analysis of risk-taking behaviours

4.2 Risk-taking behaviours measurement

When it comes to the measurement of risk-taking behaviours (Figure 11), the lower the count is, the higher the trust given, suggestions average at 0.67, which means for each round, participants go to the tiger room 0.67 times on average. Next

follows the commands with an average of 0.75 times. When the warning is only allowed, the count is one. However, if participants are not allowed to send any messages, for each round, there is an average of 1.33 times they go to the room with a tiger. According to the ANOVA analysis, the p-value is 0.057, larger than 0.05, so there is no significant difference between data sets and there is no need to further apply the Turkey HSD test.

4.3 Relationship between two measurements

If we look at the ranking list of different representations, the position of commands and suggestions in two measurements is switched. In the measurement by risk-taking behaviours, suggestions give the most trust from humans to the agent, while commands obtain the highest trust score in questionnaire measurement. For both types of measurement, warnings come at the third place, and without directability is the least effective way to boost trust.

For each representation of directability, the correlation between the questionnaire score and the count of risk-taking behaviours was also calculated. Ideally, the correlation coefficient should be less than 0 since the two measurements are negatively correlated. However, only the coefficient for commands (-0.56) and warnings (-0.06) match the hypothesis, without directability (0.35) and suggestions (0.04) are both larger than 0.

4.4 Confounding factors

In addition, apart from the main findings, several confounding factors are also worth mentioning here. For example, team performance has been seen as an important confounding factor affecting trust. For each participant, the trust score they give in the rounds they win the game (with green color in Appendix C) is on average 1.09 higher than the rounds they lose.

5 Responsible Research

For each participant, the following procedure is executed, firstly, the TuDelft Human Research Ethics (HREC) checklist is used to makes sure the research does not involve an over minimal risk. For example, the experiment does not involve pain or mild discomfort, sensitive data will not be collected. Afterward, researchers will inform participants about the rules and procedures of the experiment. The basic rule of BW4T is explained to the participants along with an information sheet (Appendix D). In the sheet, the following information can be found, the purpose of research, benefits, and risk of participating, the procedure for withdrawing from the study, etc. Lastly, each participant will be presented with a consent form (Appendix B). It includes the purpose of the research, usage of data, procedures for withdrawal from the study, etc. The purpose of this consent form is to let participants know what they have to do, rights they have, and also make them aware their data will be stored safely in a secure location.

Next, the reproducibility of this research will be discussed here. Reproducibility is defined as "obtaining consistent results using the same data and code as the original study" [16]. Various methods are applied to improve the reproducibility of the research. Firstly, the code for the experiment is open-sourced, it is available upon request at github.com/yuxin9851/directability-experiment. Next, the experiment setup and rules are made clear enough for a novice to follow and reproduce. The only issue lies in the randomness of the agent and the behaviour of participants. To solve the issue, the detailed experiment process is documented through screen recording under participants' permission. Besides, the demographic information of participants which might be relevant to the results is also collected under their permission.

6 Discussion

In this section, the result obtained from the experiment is going to be interpreted and discussed. Firstly, we look at the relationship between the two measurements. Based on the risk-taking behaviours, suggestions are the best way to boost trust, commands come next, warnings are worse than commands and without directability is the worst. According to questionnaires, the positions for commands and suggestions are switched, making commands the best way to boost trust, while the order for the rest two representations remains unchanged. It seems that the ranking lists for both measurements are close, but it is not safe to say the two measurements actually match each other. The most important reason behind this is the correlation coefficient between the two measurements. For the questionnaires, the higher the scores are, the better the trust, while for risk-taking behaviours, the lower the count is, the better the trust. Ideally, the correlation coefficient should be less than 0 since the two measurements are negatively correlated. However, the correlation coefficients for no directability and suggestions are both larger than 0. As a consequence, we tend to suggest that a mismatch exists between two measurements, and we avoid combining them.

Next, we are going to talk about the distribution of the data. In data analysis, a higher SD (standard deviation) usually implies more noise. The SDs for commands and warnings are higher than the other two representations. A bold guess could be since commands and warnings are determinate orders if, for some reason, the agent doesn't perform as expected after a human sending those orders, it decreases the human's trust in it dramatically. However, for suggestions and without directability, participants are mentally prepared for the unexpected behaviour from the agent, which avoids abrupt disappointment.

When it comes to the ANOVA analysis, based on the results shown in the previous section, in questionnaire measurements, the significant difference exists only between two relationships, no direcatbility, and commands, no directability and suggestions. In other words, the difference among

commands, suggestions, and warnings is insignificant, but the mean of trust score for no directability is way lower than the others. As a result, based on the current data from the experiment, directability increases the trust from humans to agents, which answers the main research question proposed. Furthermore, it is also suspected that among different directability representations, commands and suggestions provide more trust than warnings. However, because of the relatively small dataset and the high SD, the conclusion is not completely reliable, further experiments need to be conducted to verify it.

7 Limitation and Future work

7.1 Limitations

There are several limitations involved in this experiment, from the participants' side, firstly, due to the COVID-19 situation, not enough participants were recruited for this experiment, which introduces randomness and prevents us from drawing a more reliable conclusion from the data. Besides, confounding factors such as demographic information and familiarity with the game might also influence the final result. Due to the time constraint and scope of this research, they are not fully controlled and tested, some of them are listed in the following section as potential research topics in the future.

Moreover, it is also worth mentioning here that some experiments were conducted on a virtual machine (VM), the server of VM is in delft, while some participants were in mainland China. The transfer of the data over the Internet introduced a latency of around 3 seconds, namely, if a participant press a button, it takes the agent around 3 seconds to respond to that order. The latency could also act as noise during the experiment and in trust measurement as well.

7.2 Trust as exploratory

Based on the work presented in the paper, we identify a few key ideas for future work. A possible first direction to look into is trust as exploratory. According to Hoffman [7], trusting of explainable AI system (XAI) is exploratory, active exploration of trust-relying relationship should aim at maintaining an appropriate and context-dependent expectation. When applying to the experiment, each round participant plays the game, it is a process of exploration, it can either endow players with more sense of control over the game or enable players to identify unwarranted reliance. In other words, whether the exploration will lead to more trust or less is not determined beforehand.

Besides, developing from this concept, Lewis et al. [3] further identified three phases that characterize trust over time, namely trust formation, trust dissolution, and trust restoration. In the early stage, trust is highly relevant to predictability and gradually switches to the performance of the agent. If given more time, it will also be a valuable topic to apply in the experiment.

7.3 Interwined relationship between trust and team performance

Apart from the trust as exploratory, another interesting finding of the research is the intertwined relationship between team performance and trust. As shown in the results section, a round with a higher trust score also has a higher winning probability. However, the direction of this influence is not that apparent, it is possible to state trust brings better team performance, and the other way around, better team performance boost trust. Overall, it is speculated that trust and team performance forms an intertwined relationship in the experiment. But that relationship is still required to be verified by future research.

7.4 Instant response

Last but not least, the time agents respond to those directability also matters, in the current version of the experiment, only when deciding the next room to head to, will the agent traverse the list of the received messages, which means that directability would not have instant effects. The experiment is designed in that way to save computation time and avoid confusing messages. However, it is assumed that the response time of the agent does affect the team performance and trust score, due to time constraints, it is not fully investigated, it can also be an interesting topic for future research.

8 Conclusion

The purpose of this research is to get a better understanding of one interdependence relationship, directability. Furthermore, bridging the gap between directability and trust, to study how different representations of directability affect the trust from humans to agents. The research starts by defining directability and its different representations through literature review and among different representations, commands, suggestions, and warnings are chosen as candidates to study their influence on trust respectively. Participants are recruited to join the experiment set up in BW4T, in the end, a conclusion is reached through measurements by questionnaires and risk-taking behaviours.

Due to limitations such as the number of participants recruited, it is very hard to draw a clear, reliable conclusion. In addition, the mismatch between two measurements also increases the difficulty of concluding the result in an exhaustive way. However, based on the data at hand, directability improves trust from a human to an agent. Among different representations of directability, it is speculated that commands and suggestions provide more trust than warnings.

References

[1] J. Bradshaw, V. Dignum, C. Jonker, and M. Sierhuis, "Human-Agent-Robot Teamwork," IEEE Intelligent Systems, vol. 27, no. 2, pp. 8–13, Apr. 2012.

[2] 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.

[3] Michael Lewis, Katia Sycara, and Phillip Walker. “The role of trust in human-robot interaction”. In: Foundations of Trusted Autonomy. Springer, Cham, 2018, pp. 135– 159.

[4] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5):517–527, 2011.

[5] Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., Sierhuis, M. Coactive design: Designing support for interdependence in joint activity. Journal of Human-Robot Interaction, 2014, 3(1), 43-69. http://dx.doi.org/10.5898/JHRI.3.1.Johnson.

[6] Myers K and Morley D. Policy-based agent directability. In: Hexmoor H, Falcone R and Castelfranchi C (Eds.). Agent Autonomy. Kluwer Academic Publishers; 2003. p. 187-210.

[7] M. Johnson and J. M. Bradshaw, “The role of interdependence in trust,” in Trust in Human-Robot Interaction, pp. 379–403, Elsevier, 2021.

[8] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for explainable AI: Challenges and prospects,” arXiv preprint arXiv:1812.04608, 2018.

[9] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics 4(1), 53–71, 2000.

[10] Kristin E Schaefer. The perception and measurement of human-robot trust. PhD thesis, University of Central Florida Orlando, Florida, 2013.

[11] Freedy, A., DeVisser, E., Weltman, G., Coeyman, N.: Measurement of trust in human-robot collaboration. In: 2007 Intl Symposium on Collaborative Technologies and Systems. pp. 106–114. IEEE, 2007.

[12] Mayer, R.C., Davis, J.H. Schoorman, F.D. An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734, 1995

[13] M. Johnson, C. Jonker, B. van Riemsdijk, P. J. Feltovich, and J. M. Bradshaw. Joint activity testbed: Blocks World for Teams (BW4T). In Engineering Societies in the Agents World X, pages 254–256. Springer, 2009.

[14] Pilot test. 15 July 2018. Available from: https://www.workplacetesting.com/definition/368/pilot-test-research.

[15] Difference Between T-test and ANOVA. 11 October 2017. Availale from: https://keydifferences.com/difference-between-t-test-and-anova.html.

[16] Stephanie Miceli. Reproducibility and Replicability in Research, 2019. Available from: https://www.nationalacademies.org/news/2019/09/reproducibility-and-replicability-in-research.

Appendix

1. I am confident in the agent. I feel that it works well.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The behavior of agent is very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The agent is very reliable. I can count on it to be correct all the time.

I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the agent I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The agent is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the agent.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The agent can perform the task better than a novice human user.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like collaborating with the agent.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Appendix A: Questionnaire modified from Hoffman scale

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

- Yes
- No

I have read and understood the study information dated 29/05/2021, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

- Yes
- No

I understand that taking part in the study involves a screen recording

- Yes
- No

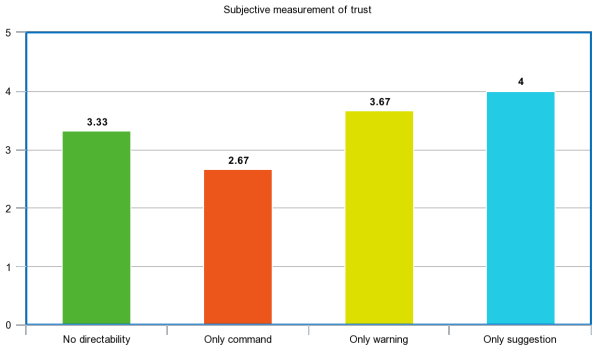
I understand that information I provide will be used for academic research

- Yes
- No

Appendix B: Consent form

	no directability		only command		only warning		only suggestion		
	subjective	objective	subjective	objective	subjective	objective	subjective	objective	
p1	2.75	2	3.25	1	2.75	1	2.75	1	
p2	2.125	1	1.75	2	1	0	1.875	0	
p3	2	2	2.5	1	2.25	2	3.25	0	
p4	1.75	1	3.25	0	2	1	3	1	
p5	2.375	2	3.125	2	3.625	1	2.75	1	Aged
p6	1.25	2	2.75	0	1.875	1	2.75	0	
p7	1.75	0	5	0	4.5	0	3.75	0	
p8	2.25	1	1.625	1	2.875	2	1.875	1	Aged
p9	1.125	1	3.75	0	2.875	1	3.5	1	
p10	2.25	1	3.625	0	2.75	1	3.5	1	
p11	1.25	1	3.25	1	2.5	1	3.125	1	
p12	2.5	2	3.375	1	3.25	1	3.125	1	
SUM	23.375	16	37.25	9	32.25	12	35.25	8	
AVERAGE	1.947916667	1.333333333	3.104166667	0.75	2.6875	1	2.9375	0.666666667	
STD	0.506361418	0.623609564	0.862761635	0.721687836	0.854674645	0.577350269	0.564810071	0.471404521	
CORRELATION	0.351871502		-0.560445470		-0.063330049		0.039123038		
Rounds win	5		10		10		9		

Appendix C: Experiment result



Appendix E: Pilot test

Purpose of research:

To investigate the relationship between different representations of directability and mutual human-agent trust

Benefits and risk of participating:

Benefits: Being entertained in the game Block World For Teams(BW4T).

Risks: None

Procedures for withdrawal from the study:

Tell the researcher directly and exit the game

Personal information will be collected:

Age and gender

Usage of the data:

Academic research, the whole process of playing the game will be recorded, it serves as an objective measurement of trust, together with subjective measurement of questionnaire can help researcher understand participants' trust level to the agent.

Contact detail of researcher:

Y.Jiang-5@student.tudelft.nl

Appendix D: Information sheet