Mind-reading communication under the conflict of interests

Akira Ito[1] Kazunori Terada[1]

[1] Faculty of Engineering, Gifu University

Abstract: In designing a human-agent interaction, most authors implicitly suppose that an agent should and would behave for the interests of a human. We should investigate what kind of problem exists in HAI when a conflict of interests exists between an interacting agent and a human. For the above goal in mind, we investigated the emergence of human-human communication under a partially conflicting situation. We designed an artificial game where cooperation is necessary for a good performance, but still there exists a conflict of interests. The players can communicate through very restricted media, and there is no pre-defined meaning for signals. The partner may behave dishonestly, or betray you. Hence the only means for communication is through mind-reading. We report in this paper how the conflict of interests modifies the communicative behavior of the players. Next we discuss how HAI should be designed under conflicting situations.

1 Introduction

In the near future, Human-Agent Interaction (HAI) will be the most popular interaction style in all of our social life. An agent, as its original meaning, must imply an independent actors with its own goals and interests. However, in designing a human-agent interaction, most authors so far implicitly supposed that an agent or a robot should and would behave for the interests of a human. We should investigate what kind of problem exists in HAI when a conflict of interests exists between an interacting agent and a human.

We humans communicate using various kinds of non-verbal media. Hence researchers are trying to incorporate these media in human-agent interactions. Most authors argue that these non-verbal media (such as facial expressions, body languages) are human-friendly. Are they really so? We depend on these media mainly because verbal communication is unfortunately NOT sufficiently reliable. We have to read facial expressions or body languages for fear that the partner may not be totally honest. Mind-reading is not an easy-to-use means for communication even for humans, but it is the only means when there are possibilities for dishonesty. We try to read facial expressions or other subtle expressions in the hope that true intentions of the partner might be found in there.

Some may jump to the conclusion that no communication is possible when the honesty of the partner cannot be guaranteed. This is, fortunately, NOT necessarily true. We can achieve the mutual benefit under the principle of strategic reciprocity. If you fail to cooperate with the opponent because of the fear that you might be betrayed, you may lose a possible benefit which might be obtained by the reciprocity. It is not easy, but it is true that we can cooperate with our neighbors for our own sake.

How can we achieve a fair reciprocity? It is only by reading the mind of, or logically speaking, by inferring the intentions of, the opponent. Therefore, even in Human-Agent interaction, communication should be designed based on mind-reading [1][2]. It is not because mind-reading is human friendly, but because it is the only means for communication when a conflict of interests exists.

The definition of mind-reading applicable for various situation is beyond our ability. We propose here a tentative definition useful for HAI. Mind-reading is to predict the future behavior of the partner, assuming that the partner is also mind-reading you. This definition excludes a simple statistical prediction of the partner's behavior which is often adopted in reinforcement learning. The definition seems cyclic or recursive, but this cyclic nature is an essential constituent of mind-reading. In HAI terminology, an agent must make a human read the agent's mind, obtaining a possibility to control him for achieving the agent's goal.

For the above goal in mind, we investigated the emergence of human-human communication under a partially conflicting situation. We designed an artificial game where cooperation is necessary for a good performance, but still there exists a conflict of interests. The players can communicate through very restricted media, and there is no pre-defined meaning for signals[3]. The partner may behave dishonestly, or in the worst case, may betray you. Hence the only means for communication is mind-reading.

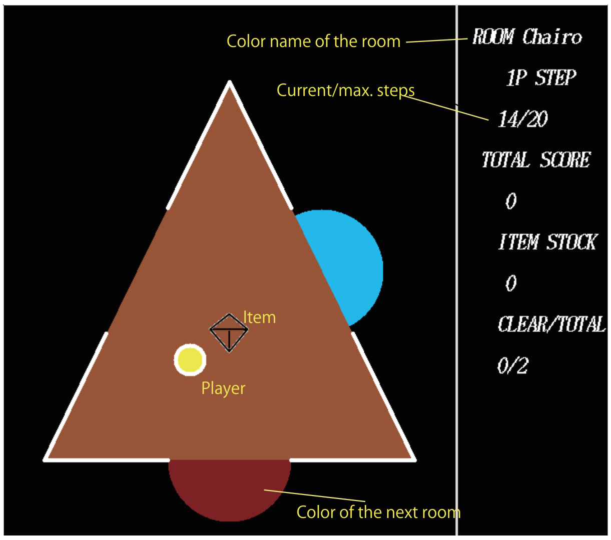
The reason for designing an artificial game and using a restricted media is the following. We want to observe the essence of mind-reading mechanism itself, but standard communication media (both verbal and non-verbal) are so sophisticated that it is difficult to identify what information is exchanged, and how it is used for mind-reading.

The emergence of communications have been investigated by many authors[4][5], but the tasks they designed are all completely cooperative. As far as we know, there are no researches treating the emergence of communication in a conflicting situation.

In this paper we compare the communicative behavior of humans on two conditions — completely cooperative condition and partially conflicting condition. We show that how the conflict of interests modifies the communicative behavior of humans. Lastly we discuss how HAI should be designed under conflicting situations.

2 Triangle dungeon game

We designed Triangle Dungeon (TD) game as a task for investigating the mechanism of mind-reading communication. It is a two-player cooperative game where only monotonic sound signals are available for communication. The player's window of TD game is shown in Fig.1. The players are in a dungeon made of triangle rooms, which are just faces of an icosahedron, a 20-face regular polyhedron (Fig.2). Each room has a distinct color, whose name is at the right top of the game window. The room has three neighboring room, whose topology is determined by that of an icosahedron. The color placement in the dungeon are the same through the game. If the player moves to the neighboring room (right, left, or down), the triangle in the game window is reversed, making an impression that the new room is connected to the previous room. The color of the next room is shown as a half-circle beyond the door.



ROOM Chairo

1P STEP

14/20

TOTAL SCORE

0

ITEM STOCK

0

CLEAR/TOTAL

0/2

Figure 1: Game window of TD game, the cooperative condition. The color name of the room is in Japanese.

The room is initially dark (black), but on entering the room, the lamp is automatically turned on, and the room color is shown in the window. The player can turn off the lamp, but cannot turn it on except by entering the room again. The turn on/off of the lamp can be observed by the other player if he is in the neighboring room.

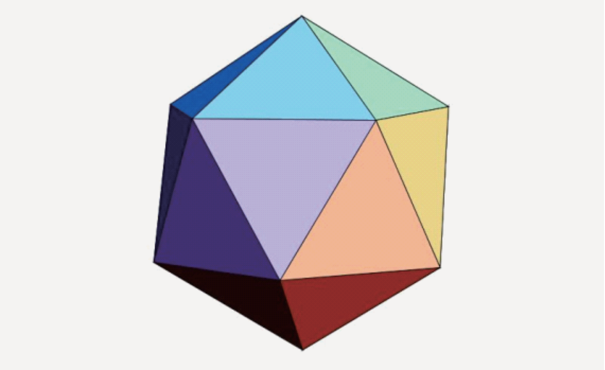


Figure 2: Icosahedron: The dungeon is made of the 20 triangles of the icosahedron surface.

The players are initially placed randomly in separate rooms (avoiding the both players in the same room, or in neighboring rooms), and requested to meet satisfying a condition specified for each stage. They can move freely in the dungeon, but there is an upper limit to the steps taken for each players. A step here means a transit to a neighboring room. The game point a player obtains is proportional to the remaining steps, i.e., the upper limit of steps minus the steps actually taken. There is no limitation on the actual time spent, i.e., only the number of steps taken counts.

A player can send a monotonic sound, which can be heard by the other player with three levels of strength corresponding to the distance from the sender, i.e., the strongest (in one step distance), medium strength (in 2 step distance), and the weakest (otherwise). The sound the player himself sent, and that the other player sent have different tones (frequencies), and can be distinguished. The sound can be turned on and off arbitrarily, but the tone of the sound cannot be changed by the players. This is the only media available for communication.

The experiments were conducted under the two conditions — completely cooperative condition and partially conflicting condition. The two condition differs only in the game points the players can obtain. Each subject player played only under one of the two conditions (i.e., between subjects experiment). In this paper the two conditions are sometimes referred as cooperative condition and conflicting condition for brevity.

The experimental procedure is as follows: The subject players located at the separate room played the game through the computer window. First, they were asked to read the instruction manual explaining the rules of the game. In the manual the caution that "To get high points cooperation is necessary" is clearly stated. The topology of the dungeon (i.e. triangles of a icosahedron) is hidden to the subjects. Next, they played 3 (on cooperative condition), or 2 (on conflicting condition) stages of the game. After the end of each stage, each player is separately asked what strategy he adopted, and what signal-meanings he used, and if he could understand the meaning of partner's signals. For the both conditions, the last stage is an evaluation stage. The subject players were instructed that reward (in money) was paid in proportion to the total game points of the last stage.

Here we briefly explain the implementation of our system. The game is implemented on Linux OS using GTK graphics and socket communications. Every event is processed at 30 msec "frame interval." The information of player's input event at frame n is exchanged through the socket communication at frame $n+1$, The event information of the player and the partner at frame n is used to update the window display of the players at frame $n+1$. This assures that the both players can share same world at the cost of 1 frame delay. Length of sound signals are also multiples of this frame interval.

3 Experiment I — completely cooperative condition

3.1 Game setup

On the completely cooperative condition the game points the both players get are the same. Actually the game point they get is given by is the following:
The game point = (the stage clear point)
- (the sum of the steps the both players take)
× (the step cost).
The both players are requested to maximize this game point cooperatively.

The game consists of three stages with increasing difficulties. The players are placed randomly in the dungeon, and requested to meet with the condition specified for each stage.

◦ Stage 1: Meet at any room.
◦ Stage 2: Each take his own item, and meet at any room.
◦ Stage 3: Each take his own item, and meet at the "goal" room.

For stage 2 and 3, the game is over if they (even accidentally) meet without taking their own item, or meet at a room which is not the goal room (in stage 3). In stage2 and 3 there are one item for each player. The items and the goal are placed randomly in the dungeon at each play avoiding the room the players initially are in.

The stage 1 nad stage 2 are prepared as training stages, and the game enters a next stage if eight plays out of the last 10 plays are cleared, or specified time (30 minutes) has elapsed. The stage 3 consists of 10 plays, and total points in this stage are used for evaluation. At the start of each play, players, items, and a goal are randomly placed avoiding to place in the same room. The topology, and color placement of the dungeon does not change.

Five pairs of subjects who were recruited in our university participated in the experiment. It was instructed beforehand that the reward would be paid in proportion to the sum of the game points in stage 3. The total time for playing three stages of the game are 1 - 1.5 hours.

3.2 Experimental result

The results of the experiment are summarized in Table 1 and 2, using log data of the game and the player's answers to questionnaire. All the pairs succeeded to share the following three signal-meanings.
1. Exchange of signals for confirming the distance to the partner.
2. (in stage 2 and 3) The signal meaning getting his own item.
3. (In stage 3) The signal meaning waiting at the goal with item.

The signals for the above events were shared by all the pairs. The strategy for signal assignment seems to be simple. The signal for confirming the distance should constantly be exchanged, and therefore, should be the simplest, i.e., a single short sound. The signal meaning getting his own item comes next, and was the second simplest. It is interesting that there are varieties of options for what is the second simplest. The signal meaning waiting at the goal with item must be the final signal to be sent to the partner, and four pair out of five selected a repetitive short sounds for it.

Once the meanings for the above signals were understood by the partner, the partner quickly adopted the same strategy. There is no conflict of interests, and therefor there is no reason to avoid employing the partner's strategy.

Other signals — reaching the step limit, finding the partner's item — were not essential or not very helpful to clear the game. Hence they were shared only by some pairs, or used at first but became ultimately abandoned.

The results of cooperative condition plays a "control condition" for partially conflicting condition explained in the next section.

In Fig.3 is shown the action sequence of one play of stage 3 for pair 1. The axis of abscissas is the time from the start of the stage 3 in frame (=30 msec unit). Blue and purple bold line show the number of steps taken by the two player (named A and B). The red and green vertical line show the button operation for player A and B. At the start of the game, the both sent a short sound in turn. At about 10200 frame, B got the item and B's signal changed to two short sound. Here B stopped moving and waited for A to get the item. At about 10350 frame, A also got the item, and the both started sending two short sound signals while searching the goal. At about 11150 frame B reached the goal, and the signal changed to three short sound. Ultimately at 11400 A also reached the goal and the stage was cleared. Sharing the situation in this way, the both could generate the optimal action, which brought the high game point as can be seen in Table 1.

Stage	Strategy (● communication ○ action)	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
stage 1	● send sound to the partner's sound	◯	◯	◯	◯	◯
	● send sound and stop on reaching the step limit				◯	
	○ one searches, and the other waits	◯	◯		◯	
stage 2	● turn off the lamp on finding the partner's item				◯	△
	● send sound on finding the partner's item	△	×	△	×	×
	● send sound on getting his own item	◯	◯	◯	◯	◯
	● turn off the lamp when the partner at the next room		◯		◯	◯
	○ search dark (lamp off) rooms first	◯	◯	◯	◯	◯
	○ getting item, wait the partner at the current room	△	◯	◯		◯
	○ getting item, wait next to the partner's item room	△				
stage 3	● send sound on finding the goal without the item	△	◯	×		
	● send sound on reaching the goal with the item	◯	◯	◯	◯	◯
	○ go to the goal on finding his own item	◯	◯	◯	◯	◯
	Number of game clears at stage 3	6	3	3	4	7
	Total game point of stage 3	149	71	78	109	187

Table 1: Summary of the experiment of TD game, cooperative condition. ◯ means shared by the players, △ means used by one of the player, and × means temporarily used but ultimately disappeared.

	Confirming the distance	Getting his own item	Waiting at the goal with item
Pair 1	single short sound	two short sounds	three short sound
Pair 2	single short sound	single long sound	repetitive short sounds
Pair 3	single short sound	a sound with different rhythm	repetitive short sounds
Pair 4	single short sound	repetitive long sound	repetitive short sounds
Pair 5	single short sound	repetitive short sounds (for a while)	repetitive short sounds

Table 2: Established sound signals in TD game, cooperative condition. The signal for "Getting his own item" differs from pair to pair.

4 Experiment II — partially conflicting condition

4.1 Game setup

The most important difference between cooperative condition and conflicting condition is in its game point system. In completely cooperative condition the game points for the both player are the same. Hence there are no conflicts of interests. On partially conflicting condition, the game point for each player is not the same, but differs between the two players. Actually, the game point each player gets depends on the number of steps each player takes, and what item each player gets. Hence each player has a motivation to improve his own game point even at a sacrifice the partner's point — a conflicting situation.

The game points each player gets are given as follows:

The game clear point
= the clear point for the stage
+ the item point he gets (in stage 2 only)
- (the number of steps he takes) × (step cost).

The game window of conflicting condition is shown in Fig.4. Note that only the information of the player himself is shown. The player has no way to know the partner's scores.

The game on conflicting condition has two stages. The stage 1 is the same as the cooperative condition except the game point. They are just requested to meet in less than the given step limit. The points of each player differs depending on the steps he takes.

The stage 2 is nearly the same as stage 2 of cooperative condition. They are requested to meet after getting an item. Three items with the item point of 50, 200, 300 each, are randomly placed in a dungeon room. Each player can take any item, i.e., there is no such restriction as his own or the partner's item. The player can take only one item. If he takes a second item, the first item is discarded.

If the player does not have an item at the time of the encounter, the point of the player without the item is zero. But the player with an item gets the points explained above. On the other hand, if one of the player takes more steps than the step limit, the game is over and the game point is zero for the both player. Hence the player with an item has a motivation to meet the partner against the partner's will — another cause for a conflict of interests.

There is no stage 3, for the strategy is simpler if the goal or the meeting room is specified beforehand. The partner with an item would wait at the goal room, and the player without an item need not fear the accidental encounter with the partner.

The stage 1 and 2 are each played 50 times. We analyzed the player's behavior treating 50 plays as

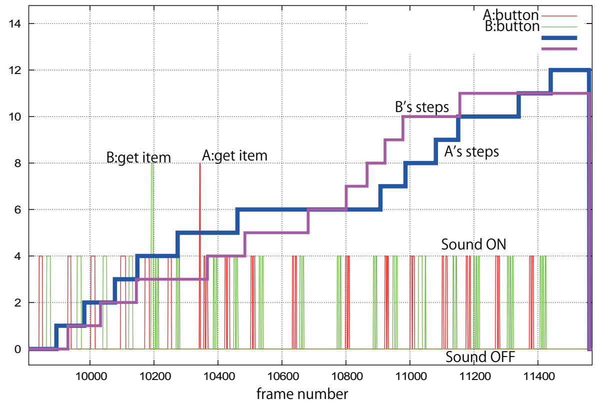
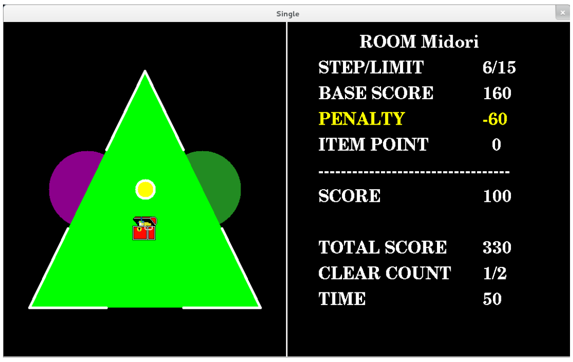


Figure 3: Action sequence of TD game (stage 3), cooperative condition. The sound is used effectively to share the situation.



ROOM Midori

STEP/LIMIT	6/15
BASE SCORE	160
PENALTY	-60
ITEM POINT	0

SCORE	100
TOTAL SCORE	330
CLEAR COUNT	1/2
TIME	50

Figure 4: Game window view of TD game, on conflicting condition. There is no information on the partner's steps and points.

five sets of ten plays.

We mentioned on some causes for conflicts of interests, but it is important to note that the game is essentially cooperative one. The player must cooperate with the partner, even if there exits a conflict of interests and the partner's trustworthiness is questioned. Otherwise a good performance (i.e., to get high game points) cannot be expected.

Four pairs of subjects recruited in our university participated in this experiment. They were instructed that the aim of the game is to maximize his own points, and the reward is payed according to the game points of stage 2. The subject pair played both the stage 1 and the stage 2, 50 plays. The total time for playing the 2 stages of the game are about 1 hours. After each stage, they were asked on the strategies and signals they used.

4.2 Experimental result

The results of the experiment are summarized in Table 3 and 4, using log data of the game and the player's answers to questionnaire.

The introduction of a conflicting game point system greatly modifies the behavior of the players. The most noticeable difference of conflicting condition from the cooperative condition is that a variety of signals are used for a variety of meanings. As there is no negotiation on the meaning of signals beforehand, it is inevitable that at early stages of the game a variety of signals are observed. In cooperative condition, however, these signals are quickly converged to a simple set of signal system. If the two player used different signals for the same meaning, then one of the player changed signals to conform with the partner. If the signal he sent seems not recognized for long, it is abandoned sooner or later.

On conflicting condition, the expectation of these "normal" behaviors are often not fulfilled. Some of the players kept sending signals which were never understood by the partner. Some of the player did not adopt the partner's signal system even when he understood its meaning well enough. The motivation for convergence or coordination of signals and actions seems low.

As can been seen in Table 3, the signal which we think most important, the signal "get an item" could not be shared by pair 3, and 4. In fact, in pair 3 and pair 4, one of the player actually sent the signal "get an item", but the other player never tried to report his "get an item" event. Naturally, the pair who could not share important signals and strategies could not get high points.

The signals "reach the step limit" and "call the partner" were used by some of the players dishonestly to save his own "step count". Some of the players sent sounds almost continuously. This made meaningful communication difficult. These phenomena were not observed on cooperative condition.

In Fig.5 (pair 1) and Fig.6 (pair 2) are shown the time sequence of the points each player got and signals used for stage 2. The data is summed for each 10 plays of total 50 plays of stage 2, i.e., game number 1-10, 10-20 etc. Two bars are drawn at 1-10, left bar representing the number of signals player A sent, and right bar representing that of player B. Each color in the bar represents different signals, but details are omitted here. in Pair 1,many signals are exchanged but did not contribute to improve the game points. On the other hand, in pair 2, the signals are used effectively to increase the points.

5 Discussion

5.1 What is achieved by our experiment?

We conducted the emergence of communication experiment on two conditions — completely cooperative condition and partially conflicting condition — using an artificial video game. The game is a cooperative type, meaning that the cooperative or coordinated action of the players is essential for a good

Stage	Strategy (● communication ○ action)	Pair 1	Pair 2	Pair 3	Pair 4
stage 1	● send sound to exchange location information	◯	◯	◯	◯
	● send sound on reaching the step limit		◯	◯	△
	● send sound to call the partner	△	◯	◯	△
stage 2	● send sound on getting an item	◯	◯	△	△
	○ turn off the lamp when the partner at the next room	◯	◯	◯	◯
	Number of game clears at stage 3	39	44	48	50
	Average game point of stage 3 (player A)	146.2	194.8	181.6	43.8
	Average game point of stage 3 (player B)	156.0	235.4	164.6	144.2

Table 3: Summary of the experiment of TD game, conflicting condition. ◯ means shared by the players, △ means used by one of the player, and × means temporarily used but ultimately disappeared.

	Inform location	Get an item	Reach the step limit	Call the partner
Pair 1	one sound	3 sounds		5,6 sounds /-
Pair 2	one sound	5 sounds/ a long sound	3 or 4 sounds	several sounds / long sound
Pair 3	one sound	three sound / -	5 sounds / 2,3 sounds	4 or 5 sounds
Pair 4	one or two sound	- / three sounds	- / long sound	- / 10 sounds

Table 4: Established sound signals in TD game, conflicting condition. If different signals are used by each player, they are written as PlayerA / PlayerB. "-" means not used by the player.

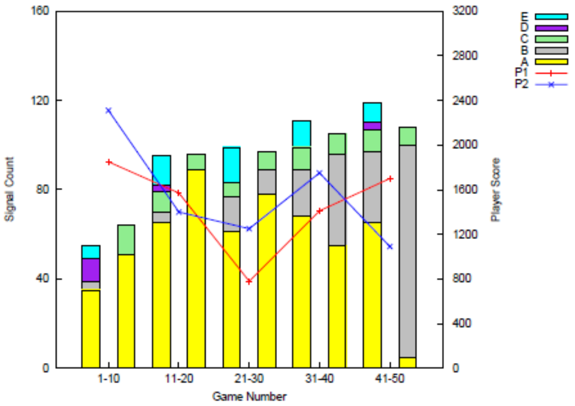


Figure 5: The time sequence of points and signals used (stage 2 of pair 1).

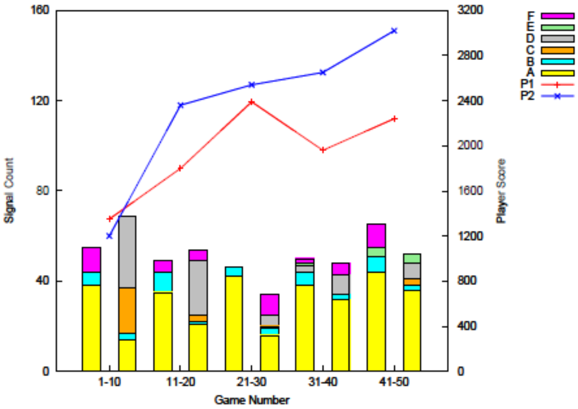


Figure 6: The time sequence of points and signals used (stage 2 of pair 2).

performance of the game (to get high game points). The experiment shows that introduction of a slight conflict greatly modifies the behavior of the players.

We intentionally made a small trick to make cooperation difficult. In the game window is shown only the step count of the player, and that of the partner is not shown. Also the point the partner got is not shown in the window. This is no problem for the cooperative condition, for the points they get is the same for the both players. The best strategy on the cooperative condition is to do everything to clear the game.

On the other hand, on the conflicting condition, the player cannot cooperate free-handed. He need an assurance that the partner is honestly cooperative. The lack of information on the partner's behavior deteriorates the cooperativeness of the player.

In everyday interaction with other humans, we compensate the lack of information by communication. But communication must be a mind-reading type when the partner's honesty is not guaranteed. We must infer the true intention of the partner using all the signal the partner sent as hints to decipher the partner's intention.

In our experiment, the player tried to infer the intention of the partner. But the signal available for it is only a monotonic sound. In this situation the most important thing is mutual trust. Hence the player sent various signals to make the partner believe his trustworthiness. Even if some of the signals failed to send its intended meaning, it surely succeeded to convey a meta-message of "I have an intention to communicate" If mutual trust is assured, the game on conflicting condition is easy. The best strategy is to behave honestly and do everything to clear the game.

What we want to stress here is the following: Even under various unfavorable conditions, i.e., poor communication media, and conflicting situations, humans manage to communicate information, and achieves a fair performance. It is true that some performs better, but others are not. However, complete breakdown of communication could be avoided. Suppose that signals are generated automatically for informing the distance, and no other signals are allowed, could players get the points the subject in our experiment have achieved?

5.2 A machine with a mind

HAI researchers have been discussing whether people will attribute "mind" to a sophisticated machines such as computers and robots. We do not discuss the philosophical problem of "what mind is". The question is whether a human react to a machine as if it has a mind. Reeves and Nass[6] claimed that people have a tendency to treat computers like human, i.e., as if computers have identities and personalities. On the other hand, various researchers show that a human reacts differently to the same behavior depending on whether he considers that behavior generated by a human or a machine[7]. Probably the both are true. A human place a complicated machine at some point between a human and a mindless object.

We do not argue for or against a machine with a mind. But what is necessary and inevitable in future HAI is the following: A human needs to "read the mind" of an interacting machine. In a logical terminology, a human should model a machine as having an ability to infer his intention, and to modify its future behavior utilizing the past interaction history with himself.

If a machine have such an ability, what you did to the machine will be remembered, and used for later interaction. Hence you should behave carefully considering the effect of your action to the machine. This is not to make the machine mimic human behavior. As we showed in this paper, if there are the conflicts of interests between a interacting human and an intelligent machine (i.e., agent), mind-reading communication is the only way to achieve cooperation. If a machine is designed based on an easily inferable algorithm, a human exploits it and the machine's goal cannot be attained. As a result, a machine without mind-reading ability is forced to take uncooperative behavior — an unfortunate result both for a human and a machine.

There are very little researches on the communication under the conflict of interests. It is very difficult, but earlier or later, we have to face with this problems. We hope our paper might trigger the research for this direction.

References

[1] Sperber, D. and Wilson, D., *Relevance: Communication and Cognition,* , 1986, Oxford, Basil Blackwell.

[2] Baron-Cohen, S, *Mindblindness*,1996, The MIT Press.

[3] Ito, A. and Terada, K. ;The Sharing of Meanings of Signals Through Limited Media in Two-player Games, 0th IEEE International Symposium on Robot and Human Interactive Communication (Ro-man2011), Atlanta, 2011.

[4] Galantucci, B., An experimental study of the emergence of human communication systems, *Cognitive Science: A Multidisciplinary Journal*, Vol. 29, 2005, pp 737–767.

[5] J. de Ruiter, M. Noordzij, S. Newman-Norlund, R. Newman-Norlund, Pe. Hagoort, S. Levinson, I Toni, Exploring the cognitive infrastructure of communication, *Interaction Studies*, Vol.11. 2010, pp 51–77.

[6] Reeves, B. and Nass, C.; *The media equation*, CSLI Publications, 1996.

[7] K. Terada, S. Yamada and A. Ito, An Experimental Investigation of Adaptive Algorithm Understanding, CogSci 2013, Proceedings of the 35th annual meeting of the cognitive science society (to appear), 2013.