Computer Poker and Imperfect Information: Papers from the 2015 AAAI Workshop

Solving Hanabi: Estimating Hands by Opponent's Actions in Cooperative Game with Incomplete Information

Hirotaka Osawa

Univeristy of Tsukuba, 1-1-1 Tenno-dai, Tsukuba, Ibaraki, Japan
osawa@iit.tsukuba.ac.jp

Abstract

A unique behavior of humans is modifying one's unobservable behavior based on the reaction of others for cooperation. We used a card game called Hanabi as an evaluation task of imitating human reflective intelligence with artificial intelligence. Hanabi is a cooperative card game with incomplete information. A player cooperates with an opponent in building several card sets constructed with the same color and ordered numbers. However, like a blind man's bluff, each player sees the cards of all other players except his/her own. Also, communication between players is restricted to information about the same numbers and colors, and the player is required to read his/his opponent's intention with the opponent's hand, estimate his/her cards with incomplete information, and play one of them for building a set. We compared human play with several simulated strategies. The results indicate that the strategy with feedbacks from simulated opponent's viewpoints achieves more score than other strategies.

Introduction of Cooperative Game

Social Intelligence - estimating an opponent's thoughts from his/her behavior – is a unique function of humans. The solving process of this social intelligence is one of the interesting challenges for both artificial intelligence (AI) and cognitive science. Bryne et al. hypothesized that the human brain increases mainly due to this type of social requirement as a evolutionary pressure (Byrne & Whiten 1989).

One of the most difficult tasks for using social intelligence is estimating one' s own unobservable information from the behavior of others and to modify one's own information. This type of reflective behavior - using other behavior as a looking glass - is both a biological and psychological task. For example, the human voice is informed by others via sound waves through the air, but informed by him/herself through bone conduction (Chen et al. 2007). In this scenario, a person cannot observe his/her own voice directly. For improving social influence from one's voice, one needs to observe others' reactions and modify his/her voice. Joseph et al. also defined such unobservable information from oneself as a "blind spot" from a psychological viewpoint (Luft & Ingham 1961).

In this study, we solved such reflective estimation tasks using a cooperative game involving incomplete information. We used a card game called Hanabi as a challenge task. Hanabi is a cooperative card game. It has three unique features for contributing to AI and multi-agent system (MAS) studies compared with other card games that have been used in AI studies. First, it is a cooperative card game and not a battle card game. Every player is required to cooperate and build a set of five different colored fireworks (Hanabi in Japanese) before the cards run out. This requires the AI program to handle cooperation of multiple agents. Second, every player can observe all other players' cards except his/her own. This does not require a coordinative leader and requires pure coordination between multiple agents. Finally, communication between players is prohibited except for restricted informing actions for a color or a number of opponent's cards. This allows the AI program to avoid handling natural language processing matters directly. Hanabi won the top German game award due to these unique features (Jahres 2013).

We created an AI program to play Hanabi with multiple strategies including simulation of opponents' viewpoints with opponents' behavior, and evaluated how this type of reflective simulation contributes to earning a high score in this game.
The paper is organized as follows. Section 2 gives background on incomplete information games involving AI and what challenges there are with Hanabi. Section 3 explains the rules of Hanabi and models. We focused on a two-player game in this paper. Section 4 explains several strategies for playing Hanabi. Section 5 evaluates these strategies and the results are discussed in Section 6. Section 7 explains the contribution of our research, limitations, and future work, and Section 8 concludes our paper.

Related Work on Incomplete Information Games

Related Trials Involving Card Games

A game-playing agent has been a challenge from the beginning of AI research (Abramson 1989). Several two-player board games with perfect information, such as Checkers, Othello, Chess, and Go, have been used as a trials by applying several new algorithms (Krawiec & Szubert 2011)(Gelly et al. 2012). In these games, all information is observable by both players. An AI system just needs to handle the condition of the board and does not need to read a cooperator’s thoughts. On the other hand, card games have unobservable information from other players (Ganzfried & Sandholm 2011). This is also an important field in AI. Poker is one of the most well-known examples, and several theoretical analyses have been conducted (Billings et al. 1998)(Billings et al. 2003). Other games including Bridge and the two-players version of Dou Zi Zhu (a popular game in China) have also been studied (Ginsberg n.d.) (Whitehouse et al. 2011).

Unique features of Hanabi

Compared with the abovementioned card games, Hanabi has three unique features that are appropriate for study in the AI field. First, Hanabi is a cooperative card game and not a battle card game. Every player in Hanabi needs to cooperate and play their own cards to create sets of fireworks before the cards run out. Second, every player can observe all of the other players’ cards, except his/her own. This means that there is no player who has an objective viewpoint. Such a condition makes a task difficult because every player needs to cooperate with others without making leaders. This is also similar to a coordination problem in multi-agent systems with incomplete information (Zlotkin & Rosenschein 1991). Finally, communication between players is prohibited, except for determining informing actions for a color or the number of cooperator’s cards. A player can only give certain information about a cooperator’s cards, e.g., the color or one of the numbers in the player’s hand. It is also possible to use other players’ plays or discards as information. This requirement avoids the difficulty of handling natural language processing. This requirement also avoids the handling of cheap talk in game theory (Wärneryd 1991). To use Hanabi as a multi-agent coordination task, we can evaluate the social ability of an AI for reading a cooperator’s intention by monitoring the cooperator’s behavior without language communication.

Definition of Restrictions

Two-player Rules of Hanbi

Hanabi is played with two to five players. In this study, we focused only on a two-player scenario.

The game is played with a deck of cards consisting of the numbers 1-5, each of which is in five different colors. Each player is dealt a hand of five cards, but the catch is that they may not look at them. Instead, the players hold the cards facing their fellow players. Players must give each other clues and use some deduction to play their cards correctly. Players will need to collectively play their cards in increasing order and by color. The game ends when the players have either completed all five sets, run out of cards in the draw deck, or made three errors during the game. They then total their score and see how close they are to a perfect score of 25.

The game comes with 50 cards that depict a fireworks explosion on the front. Each card has a number in one of five different colors. There are three number 1 cards, two number 2 to 4 cards, and one number 5 card in each color. The total cards in each color is 10 cards. The game also includes 11 tokens (eight blue tokens and three red tokens) to track both clues in the game and how many mistakes the players have made. Figure 1 shows the cards and tokens.

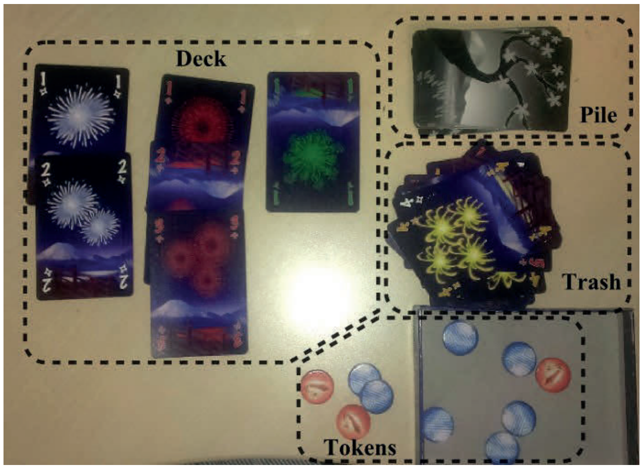

Figure 1. Components of Hanabi (in the real world).

The game play in Hanabi is simple. The 50 cards are shuffled together, and each player is dealt a hand of five cards that they may not look at. Players must hold their cards facing outward towards their fellow players. The goal will be for the players to correctly play their cards in an ascending numerical sequence. Cards must also be played according to color. One player is chosen to be the starting player, and each player takes a turn in a clockwise manner. In each turn, a player can perform one of three actions.

Give a clue; Players start with eight clock tokens. By spending one of these, they may give a clue to one of their fellow players about the cards they hold. All clues must be something that you can point to (i.e., you have one yellow card here, or these specific three cards are 4s). You cannot say "you do not have any red cards." The other caveat is that you must tell the player all of something. Thus, if you tell them that they have a red card, you must tell them each red card they hold.

Discard a Card; Once the players have used up all eight of their clue tokens, they only way to get one back is by discarding a card. A player then draws a replacement card.

Play a Card; A player chooses a card from their hand and plays it to the table. If the card is a legal play, it goes into the corresponding pile. If there is nowhere to play the card, then it is discarded, and the players lose a fuse token. If they make three such mistakes, the game ends, and the players have lost.

The game continues in this manner until either the players have successfully played all five colors of cards up to the number 5, or, more likely, the players run out of cards in the draw pile. At that point, the players each get one more turn, and the game is over. They total up the highest played card in each pile, and this gives them their total score. There is a little chart included in the game to rate how they did up to a perfect score of 25 points.

Description for Card Play

We now define the notations of several cards.

Card names

Each card is described by the first letter of the color and the number. For example, the red 4 card is denoted as R4, and the green 1 card is denoted as G1. Any card that has incomplete information, uninformed information, is denoted with an underline. For example, R_ means any red card, i.e., R1, R2, R3, R4, and R5.

Description

The board information is described as a pile P, trash T, deck D. P is a set of cards that is unobservable by both players (P = {Y3, W1, R2,…}). T is the set of discarded cards and is observable by both players (e.g., T = {R3, G2, Y1,…}). D is the set of numbers for each fireworks set (D = {W:0, B:1, Y:0, R:2, G:3}). At the start of the game, P is 50 shuffled cards, and T is zero. D is a set of zeroes (D = {W:0, B:0, Y:0, R:0, G:0}). The score is defined as the sum of D.

Each player has its own viewpoint for the world W (Wpl or Wop, where "pl" denotes the player, and "op" denotes the cooperator), which includes the card state C for each player (Cpl, Cop). C is described as several possible card sets (e.g., C = {R1, R2, W1, W2, G3} or {R1, R2, W1, W2, G4}).

Player Roles

Each player is described as a function F. The input of F is D, P, T, or W, and the output of F is defined as an action A (A = F_pl(D, P, T, W_pl)).

What is informed by informing action

If the cooperator gives an attribute of a card, the other cards' attributes are informed by this restriction. For example, if one card in the player's hand is informed as red, the possible color of the card is only red, and the other cards' possible color is any color except red.

Definition for playable card

If a card has enough information to be placed on a fireworks set, the card is defined as a playable card. For example, if there is no fireworks set, and one card is informed as "1," it is playable if the card is any color. If there is a green fireworks set, and the number 3 is on top, a player can play G4 if the player has it in their hand

Definition of discard-able card

If a card has enough information but is not playable after the turn, this card is defined as a discard-able card. For example, if there is a red fireworks set with the number 4 on top, R1, R2, R3 are discard-able cards. If there is a yellow fireworks set with the number 5 on top, any yellow card is discard-able without needing information about the number.

Definition of multiple cards

If there is a possibility that the same card is still in the deck, the card is called a multiple card. For example, if the card is R2, and there are no unobservable cards, the card is defined as a multiple card because other R2 cards may be in the pile or cooperator's deck.

Strategies

In this section, we describe five computer strategies implemented for playing Hanabi.

Ideal Strategy

In this study, we set up an ideal (cheated) strategy for Hanabi in which both players know each other's hand. With this strategy, both players can play the most optimal plays. Any other strategies cannot overcome the score gained by this strategy. An informing action is only used to skip a turn. As a result, play and discard are the only rational actions. Each player plays as follows with this strategy.

1. If the player has a playable card, it plays the card.
2. If the cooperator has a playable card, the player selects an informing action to skip a turn.
3. If there are no playable cards for both the player and cooperator, the player selects a card to be discarded using the following steps.

3.1 If the player has a discard-able card, it discards it. If not, go to the next step.
3.2 If the player has multiples of a card, it discards one of them. If not, go to the next step.
3.3 If the player has no card for discarding, it discards the highest numbered card in his/her hand.

The viewpoints of the player and cooperator are shown in Fig. 2.

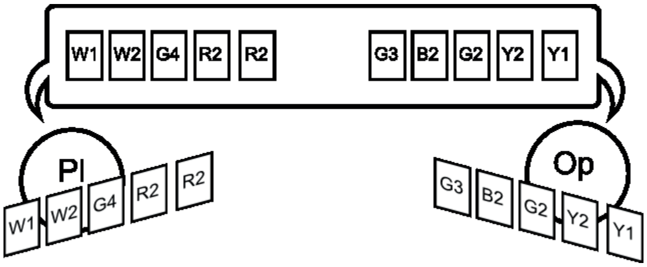

Figure 2. Ideal strategy's viewpoint.

Random strategy

In the random strategy, the player has no information about its own cards. The player randomly selects informing (30%), discarding (40%), and playing (30%) actions. We selected each of these possibilities through several simulations to maximize the results of the random strategy. The viewpoint of the player is shown in Fig. 3.

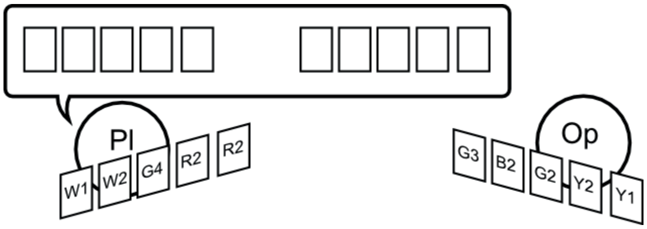

Figure 3. Random strategy's viewpoint.

Internal-State Strategy

In the internal-state strategy, each player's hand is recorded. The player uses the internal-state strategy as follows.

1. If the player has a playable card, it plays the card.
2. If the player has a discard-able card, it discards the card.
3. If the cooperator has a playable card, and there is an information token, the player states one of the attributes (color or number) of the card.
4. If there is an information token, the player states one of the attributes (color or number) of a card randomly selected from the cooperator's hand.
5. The player randomly selects one of its cards and discards it.

The viewpoint of the player is shown in Fig. 4.

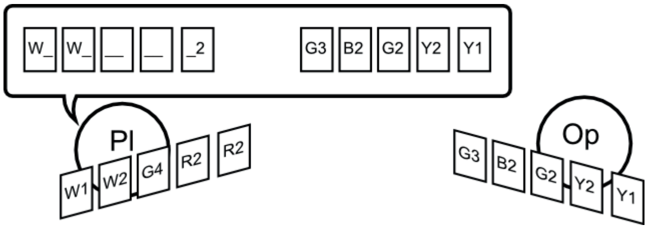

Figure 4. Internal-state strategy's viewpoint.

Outer-State Strategy

With this strategy, the player's action is similar to the internal-state strategy, but it can remember the cooperator's cards. If the player chooses to inform the cooperator about an attribute of a card (steps 3 and 4 in the previous subsection), the player selects information that has not yet been stated. The other conditions are the same as the above internal strategy. The viewpoint of player is shown in Fig. 5.

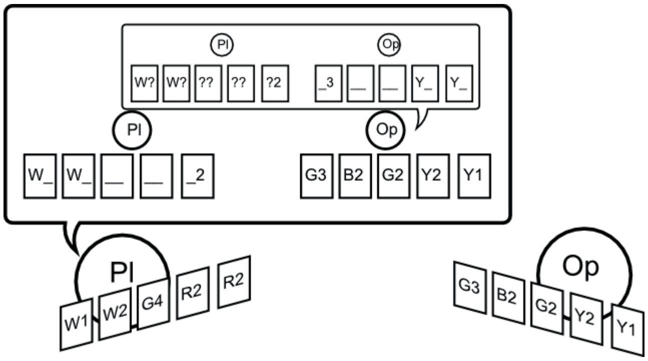

Figure 5. Outer-state strategy's viewpoint.

Self-recognition Strategy

In this strategy, each player's behavior is similar to that of the previous subsection's strategy, except one player estimates the other's intention and simulates it before step 5. The estimate of the cooperator's intention is the simulated cooperator's viewpoint of the previous turn using all possible combinations of cards induced by the player's estimation. This simulation is conducted on the premise that the player and cooperator have the same strategy. The details of the simulation process are as follows.

1. The player generates every possible combination of each hand (e.g., H = {{R1, R1, G2, G2, W1}, {R1, R1, G2, G2, W2}⋯}).
2. The player creates the previous state of the board (Dpre, Ppre, Tpre, where "pre" denotes the previous state) and the cooperator's hypothetical world viewpoint of the previous turn (Whyp_op, where "hyp" denotes the hypothetical state) according to each hypothetical possible combination.

3. The cooperator's action is simulated by the player's algorithm with a hypothetical input (Ahyp = Fpl(Dpre, Ppre, Tpre, Whyp_op)).
4. If the cooperator's hypothetical action Ahyp does not match the real cooperator's action Areal (where "real" denotes the real state), the hypothetical combination is removed from the hand.
5. Step 2 is repeated until no tested combinations remain.

According to the above process, we obtain possible player hands. After the process, we select the most probable card x and the second-most probable card y. If the possibility of x divided by the possibility of y is greater than a, we estimate that the player's hand is x. Then, we repeat the outer-state strategy.

For example, if there is no fireworks set, and the cooperator informs the player that its right-side card is green, the player simulates the cooperator's behavior on the basis of its own strategy. The player then determines if the card is playable, e.g., the card is informed as green. Then, the player estimates that it is a playable card (G1) and plays it.

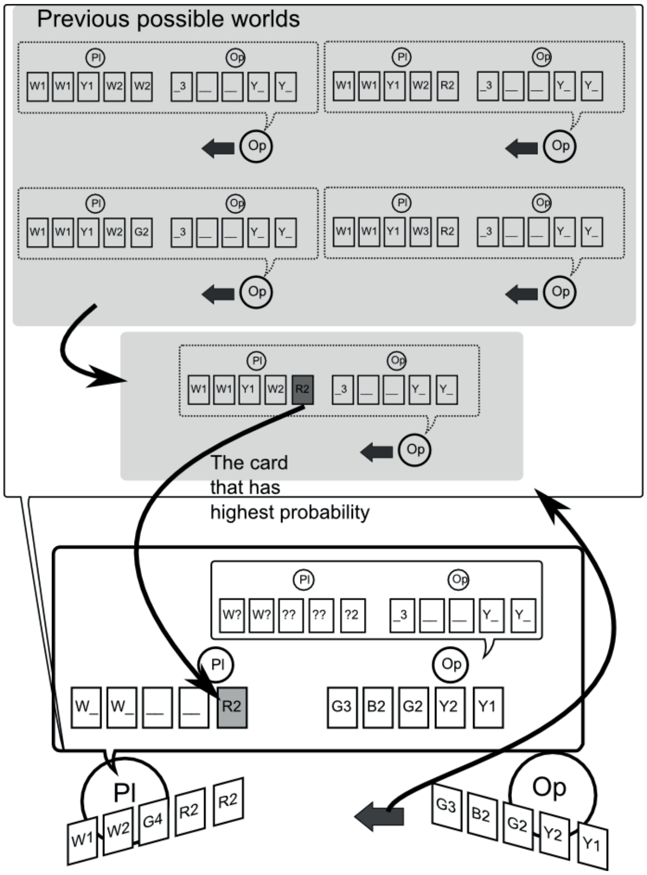

Figure 6. Self-recognition strategy's viewpoint.

Evaluation

Simulation Setup

We conducted computer simulations of players with the above five strategies. A pair with each strategy plays a game 100 times in a simulation. We set each player's hand as five cards. Owing to the limitations of computational resources and to prevent the generation of a large amount of data, we applied just one recursive simulation for the self-recognition strategy. In one recursive simulation, each agent estimates its cards from a simulation of cooperator's previous worlds. However, the hypothetical cooperator does not infer the hypothetical player's two-step previous worlds.

A threshold needs to be determined for the self-recognition strategy. Before conducting the simulation, we simulated values for a from 1 to 5 on 0.1 intervals. We found that the score is maximized when a = 2.5. We used this value for the simulation of the self-recognition strategy.

The hypothesis of the simulation is as follows. If the self-recognition strategy influences the coordination of agents, its score will be higher than the other strategies.

Result

The simulation results are shown in Fig. 7. The average score was 24.6 (SD 1.10) with the ideal strategy, 2.20 (SD 1.60) with the random strategy, 10.97 (SD 1.94) with the internal-state strategy, 14.53 (SD 2.24) with the outer-state strategy, and 15.85 (SD 2.26) with the self-recognition strategy.

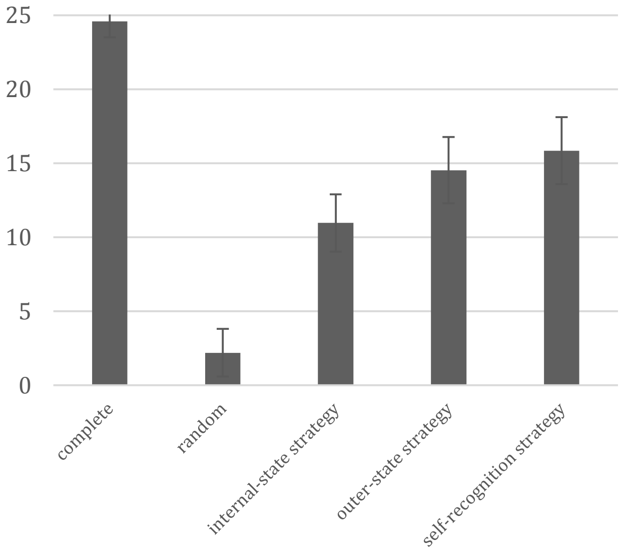

Figure 7. Scores of strategies.

We applied a statistical test ANOVA for both hands and found that all two pairs of strategies had significant differences (p < 0.01). These results suggest that the self-recognition strategy resulted in the third-best score compared with the ideal strategy (the best) and human (the second best).

Discussion

Our results support the hypothesis stated in section 5.6 and suggest that the estimation of one's hand by the self-recognition strategy is better than the internal-state and outer-state strategies.

We discuss an example of a match using the self-recognition strategy and evaluate how coordination occurs in the strategy. When D = {W:5, R:5, B:0, Y:3, G:2}, Wpre_pl = {Cpre_pl = (__,_1,__,Y_,__), and Cpre_op = (Y1, R4, Y4, B2, Y4)} and Wpre_op = {Cpre_pl=(Y3, B1, G1, Y2, B1) and Cpre_op = (__, __, _4, __, _4)}, the player told the cooperator that the number of cooperator's first card is 1, and the state changed to (Cop = (_1,__,_4,__,_4)). Although this is not deterministic information that the hand is playable, the cooperator estimated with the strategy that Y1 is the most possible card about the first card if the player thought as the cooperator thought. Then, the coopeartor estimates the card _1 as Y1 and plays card Y1. Another example is when D = {W:4, R:3, B:2, Y:1, G:2}, Wpre_pl = {Cpre_pl = (__, __, __, __, R4), and Cpre_op = (Y5, W5, Y1, R1, G4)} and Wpre_op = {Cpre_pl=(W2, B1, B3, Y1, R4) and Cpre_op = (Y_, __, __, __, __)}, the player stated that the cooperator's second card is white, and the state changed to (Cop = (Y_,W_,__,__,__)). This also did not explicitly state that the second card is playable, but the cooperator determined that this white is informed by player because this is playable. As a result, this card is played. These are the examples of complementation in incomplete information using an inference of the cooperator's intention.

The increase in score was better in the two-card-hand scenario. We believe that this is because our estimation algorithm is very simple. If there are more hands for each player, the hit rate decreases. However, if we used more modern methods such as Bayesian reasoning, we could have obtained a more precise estimation from the player's own hands. The important contribution of our work is that the player earned a better score in the simulation of the cooperator's behavior than with the other rational strategies, even with very simple estimation.

The important point is that the appropriate action automatically emerges according to the change in context without implementing heuristics in the strategy innately. We found that there are several heuristic strategies that are acquired between players by interviewing human Hanabi players. There are several tips provided by Hanabi players. For example, if there is only a red fireworks set, and the cooperator has {Y1, G1, R2, W3, G3} cards, the player informs the cooperator about the 1 cards. This action suggests that they are playable. On the other hand, if there are four 1 cards except blue, the information about these 1 cards suggests that they are not playable (and they are discard-able). If there are one blue and two white fireworks sets, the player informs the cooperator of the white card in its hand. This suggests that the card is playable. These heuristics are close to the reasoning process in the computer strategy described in the previous paragraphs. However, the simulation of the cooperator can automatically find the same solution without requiring heuristics and achieves the coordination of agents. These results indicate the importance of the theory of mind (reasoning about the cooperator's intention from the behavior of the cooperator) for general solutions studied in cognitive science (Hiatt & Trafton 2010)(Frank & Goodman 2012).

Contribution, Limitations, and Future Work

One of our contributions is that we shows the importance of inference for the cooperator's viewpoints in multi-agent coordination. Our contribution is not limited to only playing a specific card game. Our results suggest that social intelligence involving an estimation of one's own unobservable information on the basis of the cooperator's behavior becomes critical to increase the reward of the players.

However, our simulation still involved a very limited communication game. Because Hanabi has smaller branches for possible worlds, we can estimate every possible viewpoint of the cooperator. These results cannot be directly applied to other communication games because other games require communication, and this kind of free communication greatly increases the search space for branches.

Although self-recognition is effective, our estimation method from several viewpoints is relatively simple compared with our thorough simulation for viewpoints. We can apply a more sophisticated method for improving accuracy in the future. In addition, our simulation is just one recursive simulation. We removed the estimation of the hypothetical player's viewpoint in the estimation for the hypothetical cooperator. This type of recursive inference is important in joint-attention and human–agent interaction studies (Yamaoka et al. 2009). In future work, we plan to evaluate how this type of multiple recursion improves performance. Further, this simulation only uses one previous round for inference. The use of more previous rounds as a clue for inference greatly increases the number of trees. Feedback from human players suggests that complex heuristics are used in some cases. This may be the reason for the increased score of a human compared with our simulated strategy. In addition, multiple Hanabi player (3 to 5) games require more inferences about the other players. In such a situation, a Monte Carlo method may become a viable solution (Whitehouse et al. 2011).

The important findings from our results are that theory of mind and self-recognition are more general intelligence methods for meeting multiple requirements than just learning several heuristics. If we only share the other agent's algorithm and their behavioral results, we can estimate the viewpoints of the other players. This is a more general method for different cooperating agents without sharing heuristics. This may indicate why we can communicate with others sometimes without sharing context. We are all humans and can simulate other humans.

Conclusion

We used a card game Hanabi as an evaluation task of imitating human reflective intelligence with artificial intelligence. We compared human play with complete strategy, the random strategy, rational strategy without opponent's viewpoint, rational strategy with opponent's memory, and rational strategy with feedbacks from simulated opponent's viewpoints. The results indicate that the strategy with feedbacks achieves more score than only a rational strategy.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 25700024.

References

Abramson, B., 1989. Control strategies for two-player games. *ACM Computing Surveys*, 21(2), pp.137–161.

Billings, D. et al., 2003. Approximating Game-Theoretic Optimal Strategies for Full-scale Poker. In *International Joint Conference on Artificial Intelligence*. pp. 661–668.

Billings, D. et al., 1998. Opponent Modeling in Poker. In *AAAI Conference on Artificial Intelligence*. pp. 493–499.

Byrne, R.W. & Whiten, A., 1989. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*, Oxford University Press, USA.

Chen, Z. et al., 2007. Biomimetic modeling and three-dimension reconstruction of the artificial bone. *Computer methods and programs in biomedicine*, 88(2), pp.123–30.

Frank, M.C. & Goodman, N.D., 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084), p.998.

Ganzfried, S. & Sandholm, T., 2011. Game theory-based opponent modeling in large imperfect-information games. In *International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 533–540.

Gelly, S. et al., 2012. The grand challenge of computer Go. *Communications of the ACM*, 55(3), p.106.

Ginsberg, M.L., GIB: Imperfect Information in a Computationally Challenging Game.

Hiatt, L.M. & Trafton, J.G., 2010. A Cognitive Model of Theory of Mind. In *International Conference on Cognitive Modeling*. pp. 91–96.

Jahres, S. des, 2013. Spiel des Jahres Award. *Spiel des Jahres*. Available at: http://www.spieldesjahres.de/cms/front_content.php?idcatart=1121&id=828

Krawiec, K. & Szubert, M.G., 2011. Learning n-tuple networks for othello by coevolutionary gradient search. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation - GECCO '11*. New York, New York, USA: ACM Press, pp. 355--362.

Luft, J. & Ingham, H., 1961. The Johari Window: a graphic model of awareness in interpersonal relations. *Human relations training news*, 5(9), pp.6–7.

Wärneryd, K., 1991. Evolutionary stability in unanimity games with cheap talk. *Economics Letters*, 36(4), pp.375–378.

Whitehouse, D., Powley, E.J. & Cowling, P.I., 2011. Determinization and information set Monte Carlo Tree Search for the card game Dou Di Zhu. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*. Ieee, pp. 87–94.

Yamaoka, F. et al., 2009. Developing a model of robot behavior to identify and appropriately respond to implicit attention-shifting. *Proceedings of the 4th ACMIEEE international conference on Human robot interaction HRI 09*, pp.133–140.

Zlotkin, G. & Rosenschein, J.S., 1991. Incomplete information and deception in multi-agent negotiation. In *Proceedings of the 12th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 225–231.