Constructing a Human-like agent for the Werewolf Game using a psychological model based multiple perspectives

Noritsugu Nakamura*, Michimasa Inaba*, Kenichi Takahashi*, Fujio Toriumi†, Hirotaka Osawa‡,
Daisuke Katagami§ and Kousuke Shinoda¶
*Hiroshima City University
3-4-1 Ozukahigashi, Asaminami-ku, Hiroshima, 731-3166 Japan
Email: nakamura@cm.info.hiroshima-cu.ac.jp, inaba@hiroshima-cu.ac.jp
†University of Tokyo, Japan
‡University of Tsukuba, Japan
§Tokyo Politechnic University, Japan
¶University of Electro Communications, Japan

***Abstract*—In this paper, we focus on the Werewolf Game. The Werewolf Game is an advanced communication-game in which winning or losing is directly linked to one's success or failure in communication. Therefore, we expect exponential developments in artificial intelligence by studying the Werewolf Game. In this current study, we propose a psychological model that considers multiple perspectives to model the play of a human such as inferring the intention of the other side. As one of the psychological models, we constructed a "one's self model" that models the role of others as viewed from their own viewpoint. In addition, to determine whether one's opinion is reliable after inferring other's intentions, we also constructed an "others model" that models the role of others as viewed from their viewpoints. Combining these models, we showed through experimentation that a combined approach achieved better results, i.e., higher win percentages.**

I. Introduction

In this paper, we focus on the Werewolf Game, a cornerstone of communication games. It is difficult to use common game artificial intelligence (AI) methods, such as tree search algorithms because the Werewolf Game is played by conversing with a werewolf agent, i.e., AIwolf. In the Werewolf Game, it is necessary to model and describe communication that a computer can easily process. However, in the process of communication via a computer, there are technical challenges, because it is difficult to formulate a game situation.

Overcoming these technical challenges is indispensable to developing future techniques in AI, and there are multiple possibilities for contributing advances in the study of the Werewolf Game. Therefore, we focus our efforts on the subject matter of the Werewolf Game.

One of the challenges in realizing natural communication with a computer lies in the fact that often the success or failure of the communication is unclear. Therefore, s computer cannot efficiently learn and evaluation is also difficult. Conversely, it is possible to measure the success and failure of communication by leveraging winning or losing in the Werewolf Game, because winning or losing is directly linked to the success or failure of communication. Therefore, we can evaluate and learn via enforced learning, a technique that has already achieved success in game AI. Furthermore, we expect exponential development in AI, because communication in the Werewolf Game contains highly complex subject matter such as inference and persuasion and false swearing.

In our previous study, we studied protocol design for enabling AIwolf to have a conversation [1] and then released a construction kit for developing agents that play the Werewolf Game using our developed protocol [2]. In these studies, we attempted to solve the challenges for realizing sophisticated communication by improving upon an environment in which the computer play the Werewolf Game and obtain collective intelligence via a contest. A contest involving the Werewolf Game was recently held for the first time as part of the Computer Entertainment Developers Conference (CEDEC2015) in August 2015 in Japan, with over 50 teams participating.

In many game AI studies, initial stages typically succeed by constructing a model that imitates the play of humans. In this current study, we propose a psychological model that considers multiple perspectives to model the play of a human, experimentally validating our proposed model experimentally. In our experiments, we implemented the proposed model using the aforementioned construction kit for developing agents [2], and the evaluated it by setting our implementation against agents that successfully made it to the final game of CEDEC2015 We present a summary of our psychological model that considers a variety of perspectives in Fig.1. In the Werewolf Game, it is necessary to consider the other players viewpoint and infer their intention because there is little objective information other than conversation. For example, in the figure, *A* states, "I think *C* is a liar" in a situation in which *C* is the only liar among *A*, *B*, and *C*; however, we can consider multiple interpretations because *B* does not know whether *A* or *C* is a liar. In one situation, *A* is not a liar, and *A* suspects that *B* is

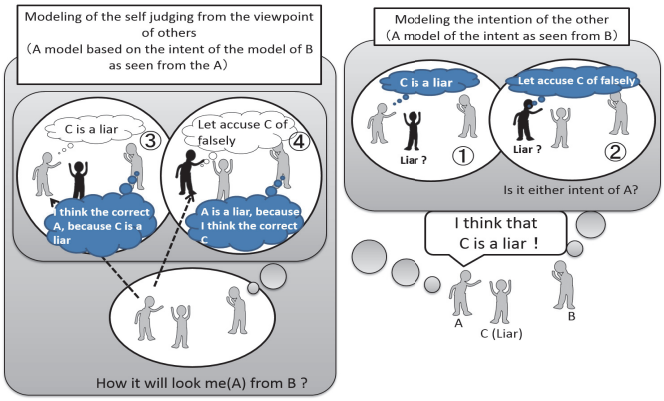

Fig. 1. Psychological model that considers multiple perspectives to model the play of human

lying (i.e., Fig.1 ①). In another situation, *A* is a liar, and *A* falsely accuses *B* of being a liar (i.e, Fig.1 ②).

The above cases comprise modeling the intention of another to attempt to identify how to successfully come to a compromise. The international Automated Negotiating Agents Competition (ANAC) [1] is held annually; however, it is insufficient to only model the intention of the other in the Werewolf Game, because it is necessary to assert that one's opinion is reliable after inferring the intention of the other, persuading a third individual in the Werewolf Game.

Therefore, modeling of one's self from the viewpoint of others (i.e., the left-hand side of Fig.1) is required. A player earns trust by not only inferring the intention of another, but also modeling one's self from the viewpoint of others. Given this, the player then reacts after considering how one looks and must persuade the other. In the example shown in the figure, *A* needs to reply after considering how *B* and *C* think of *A* when *A* states, "I think *C* is a liar." If *B* thinks that *C* is a liar (i.e., Fig.1 ③), it is expected that *B* trusts *A* since *A* has the same opinion. Conversely, if *B* trusts *C* (i.e., Fig.1 ④), *A* must consider that *B* may suspect *A*, who thinks *C* is a liar. These approaches are indispensable when humans play the Werewolf Game; thus in this study, we model these approaches and implement them in the AIwolf agent.

II. Related Work

Our psychological model adopts intention estimation based on others models [3]. As humans, we communicate smoothly by estimating the psychology of the other via various clues, taking appropriate actions with that estimated psychology in mind. It is said that not only humans, but also primates and crows can estimate the intentions of others [4], [5]. Furthermore, cooperation actions [6] and instructions to others [7] are enabled by estimating the psychology of others.

Applications of the others model in computational intelligence exist in the field of human-computer interaction (HCT). Example studies include the analysis of interpersonal interactions using simulations [8] and the construction of a physical communication model of a robot [9].

Regarding research of games that include elements of communication, there are studies focused on a Monopoly [10] and The Settlers of Catan [11]; however, there are few elements such as trust and persuasion and the flexibility of communication is low. In addition, it is possible to play these games without using communication. Conversely, the Werewolf Game has very few limits on communication, and the quality of communication is directly linked to winning or losing. Finally, there is also a study that analyzes The Resistance game, which has a lot in common with the rules of the Werewolf Game; in The Resistance, the goal is to find the spy lurking among the players [12]; however, conversation between players is not subject to this study.

Studies of the Werewolf Game are few and far between, but do exist. From studies by Braveman, et al, [13] and Yao [14], let n be the number of players at the time a game starts, and let m be the number of werewolf players. The authors of these two studies showed that the probability $w(n, m)$ that a werewolf wins is proportional to $m/\sqrt{n}$. Next, Migdal showed a rigorous formula of the probability $w(n, m)$ that the werewolf wins [15]. In these studies, it was assumed that the "seer" role, which has special abilities, does not exist; furthermore, for simplification, the game was played using only the werewolf and villager players. In addition, each player was assumed to act at random, reducing the influence of communication. There is study regarding the analysis of the strategy of the werewolf game when limited the behavior of werewolf [16]. In this study, each of the werewolf-side players behaves like a villager, and the player does not pretend to have a other role. Studies that focused on human actions and the psychological side when playing the Werewolf Game have been reported. For example, studies have focused on the number of times a specific utterance was made and the number of times that cut in on the conversation of each player [17], the movement of one's hands and head [18], and the length of the conversation [19]. Although various analyses have been performed with respect to the Werewolf Game, to our knowledge, there are no studies that constructed a player's action model in the Werewolf Game.

III. The Werewolf Game

The Werewolf Game is a party game that is played by using the advanced communication abilities that we have as humans. When a game starts, all players are randomly divided into either the villager side or the werewolf side, and then allocated to roles. We summarize each role in Table I. Players in the villager side and possessed players cannot know which side other players belong to; however, werewolf players can know friendly players.

In the Werewolf Game, there are two phases, i.e., day and night. In the day phase, all players discuss who the werewolves are, and then select a player who is targeted for execution based on a vote. In the night phase, werewolf players attack

[1] http://mmi.tudelft.nl/anac

TABLE I
THE ROLES OF PLAYERS IN THE WEREWOLF GAME

Role	Explanation
Villager	A human that does not have a special ability
Seer	In the night phase, a seer can select one player and know whether the selected player is a werewolf or not.
Medium	A medium can know whether a player who is executed by a vote is a werewolf or not.
Bodyguard	In the night phase, a bodyguard can select one player and guard the selected player from attack by a werewolf.
Possessed	A possessed player is a human in the werewolf camp, but his or her objective is the same as that of the werewolf players. In addition, a possessed player is judged as human by both a seer and a medium. If the werewolf camp wins, possessed player also win.
Werewolf	A werewolf changes oneself into a human in the day phase and can participate in the discussion. In the night phase, werewolf players attack one human player.

a human player selected based on discussions with friendly players. Moreover, players with special abilities (e.g., seer and medium) can use their special abilities at night. The executed player and the attacked player are eliminated from the game, and are not allowed to participate in further discussion or votes. Those player roles are not revealed until the game is over. All players in the game can discuss freely, and a crucial aspect for players who belong to the villager side is to detect the lies put presented by werewolf players based on discussions and special abilities. For werewolf players, the crucial aspect is to manipulate discussions to their advantage by impersonating a role.

The villager side wins when all werewolves are killed, whereas the werewolf side wins when they kill enough humans to bring the number of humans to be less than the number of werewolves. Day and night phases repeat until one group meets the respective conditions for winning. A possessed player is counted as a human, but belongs to the werewolf side.

In this paper, we model the game such that it is played by 15 players with six roles; more specifically, we have eight villager players, one seer player, one medium player, one bodyguard player, one possessed player, and three werewolf players.

IV. CONSTRUCTION OF THE PSYCHOLOGICAL MODEL

A. An overview of the psychological model

The psychological model that we propose is constructed via a psychological table that expresses which roles a certain player believes other players to be based on probabilities. In this psychological table, each utterance and action has a corresponding score that suggests that the player that made the utterance and action plays a specific role. Scores are added to each player to ultimately identify a target.

In the model, we calculate the probability using the score. In our model, probability is calculated by normalizing each role's score by summing all role scores for each player; however, the bodyguard is included in the villager count, because it is rare that a bodyguard expresses his or her own role, and there are very few specific utterance and actions. We show an example of the psychological table in Table II, with numbers in the table corresponding to the relative probabilities and numbers in parentheses expressing the corresponding scores. Our model includes the psychological model of others judging themselves (i.e., the one's self model), combining this with the psychological model of others judging others (i.e., the others model). We explained each model in the subsections that follow .

TABLE II
EXAMPLE PSYCHOLOGICAL TABLE MAPPING PLAYERS TO ROLES WITH CORRESPONDING PROBABILITIES

Player	Vill	Seer	Med	Poss	Wolf
A	0.00 (0)	0.80 (80)	0.00 (0)	0.10 (10)	0.10 (10)
B	0.70 (70)	0.00 (0)	0.00 (0)	0.10 (10)	0.20 (20)
…	…	…	…	…	…
N	0.00 (0)	0.00 (0)	0.00 (0)	0.40 (40)	0.60 (60)

B. The psychological one's self model

The one's self model models which roles one's self thinks of other players, and therefore has one psychological table. When one's role is a werewolf, the werewolf score in the table corresponding to friendly werewolves is set to 100, whereas other scores are set to zero. Essentially, the one's self model has only one psychological table. However, if one's own role is the werewolf or the possessed, it is necessary to hide the role. Therefore, the player has some psychological table of when it is assumed one's own as other roles, i.e., a villager, a seer, and a medium. Here, friendly werewolves are not considered.

C. The psychological others model

The others model models which roles other players think of other players. Here, the psychological table supposes each other player to be a villager, seer, medium, possessed or werewolf, because one does not know the actual role of others. Furthermore, when one's role is a seer, medium or possessed, the table is unnecessary, because there is not the role to other players In addition, when one supposes another player to be a werewolf, it is necessary to suppose two players are friendly werewolves. Therefore, we make a psychological table of all patterns of the three werewolves. The number of psychological tables that suppose werewolves is ${}_{13}\mathrm{C}_2 * 14 = 1092$, because we select one player from among the players but not one's self, and select two friendly werewolves from among the players but not one's self and one player that is chose earlier. This psychological table, which is unnecessary, is deleted while playing the game. The condition for being deleted is as follows:

- If all players included in the pattern have died
- If a player who is successfully identified as a werewolf is not included in the pattern
- If a player who is successfully identified as a human is included in the pattern

D. Confidence rating of the psychological table in the others model

1) Confidence rating of the psychological table in the others model in which we suppose that other players are not werewolves:

In the others model, there are plural psychological tables, since we consider the possibility of all roles of each player; however, the most important table among these is the psychological table that is consistent with the actual role of the player. In addition, when we refer to the probability from the psychological table of the others model, we must evaluate whether the probability can be how much reliable. For example, when we refer to the probability from the psychological table of the others model of a player who expresses the behavior of a seer, there is a very low possibility that the player is a villager or a medium. Therefore, there is no need to emphasize the probability in the psychological table. For these reasons, we must evaluate the psychological table in the others model. When we suppose that a certain player is a villager, seer, medium or possessed, we use the confidence rating from the psychological table based on the probability of the role of the player from the one's self model.

2) Confidence rating of the psychological table in the others model in which we suppose that other players are werewolves:

When we suppose that other players are werewolves, the psychological tables of all patterns of three werewolves are made. Therefore, the confidence of the psychological table of all patterns is required. To calculate the confidence rating of the psychological table that corresponds to the pattern of a werewolf, we obtain appropriate scores from the psychological tables of the patterns of a werewolf. Here, the term "confidence rating" indicates how probable it is that the werewolves are consistent with the pattern. Regarding utterances and actions, the degree to which both the player that uttered (i.e., acted) and the player that is the target of the utterance (i.e., action) is a werewolf were scored between zero and 100, with 50 being neutral; we add this confidence rating in the corresponding psychological table. The confidence rating of the psychological table that corresponds to the werewolf pattern is a probability that is normalized based on the total score in the psychological table of all werewolf patterns. For example, if both players *A* and *B* have expressed behavior suggesting that they are seers, a low score is set, because the possibility that both *A* and *B* are werewolves is low. When we add the scores, if the sum is 50 or more, the score is added to the confidence rating of the psychological table of werewolf patterns which includes both players *A* and *B*. If the score is less than 50, we subtract it from 50 and add this difference to the confidence rating in the psychological table for werewolf patterns that do not include either *A* or *B* or both *A* and *B*. By using this confidence rating, it is possible to consider various combinations, such as player *B* is also a werewolf when *A* is a werewolf.

E. Questionnaire for the psychological model

In the psychological model described so far, we add the specific score depending on the situation of the game. In this study, in order to realize the human-like behavior, we identify scores by a questionnaire. In addition, in the progression of the game, there is possible to express own role. The timing of express own role also depends on the situation of the game. Therefore, the timing is also determined by the questionnaire. The subjects having experience of the Werewolf Game, were recruited by a crowdsourcing site, Crowdworks[2].

1) Questionnaire regarding adding scores to the psychological table: In the questionnaire regarding adding scores to the psychological table, when the given situation, we asked participants to assign a score to each of the roles of the specified player in the 10 intervals of zero through 100. The average of these scores is then added to the psychological table.

2) Questionnaire regarding the timing and expression of roles:

We used this questionnaire to determine the timings that express certain roles. The timing expresses the role of seer or medium is as follows:

- If all players have not expressed behavior for their role
- If other players have expressed their respective roles
- If the seer or medium uses his or her special abilities to identify a werewolf

. For these situations, we created a question that asked what the probability was for a player to express behavior of a seer or medium. For possessed and werewolf players, we specified the situation of the given role being expressed, and at the time, created questions that asked for probabilities of one claiming to be a seer or medium.

3) Questionnaire regarding the werewolf pattern in the others model:

For this questionnaire, we created questions that asked the possibility that a specific pair of players were both werewolves. For example, "If both players *A* and *B* have expressed behavior suggesting that they are seers, how likely do you think it is that both players *A* and *B* are werewolves? ". We created 15 questions about expressing roles, results of divinations, results of inquests, votes, and estimating roles. By using these scores, the confidence values shown in Section IV-D2 are set.

V. Action decisions of an agent

In this section, we describe action decisions of an agent that uses a psychological model. The agent was created using the our AI platform for the Werewolf Game, which is a platform for a simplified Werewolf Game that can perform the utterances shown in Table III, vote, divination, inquest, guard, and attack.

[2]https://crowdworks.jp/

TABLE III
FUNCTIONS AND UTTERANCES THAT THE AGENT CAN PERFORM

Estimate(AGENT, ROLE) I think that the AGENT is ROLE	Comingout(AGENT, ROLE) AGENT expresses ROLE
Divined(AGENT, SPECIES) When a divination occurs the result is that the AGENT is SPECIES	Inquested(AGENT, SPECIES) When an inquest occurs, the result is that the AGENT is SPECIES
Guarded(AGENT) I guarded AGENT	Vote(AGENT) I voted for AGENT
Agree(TalkType, day, id) I agree with the utterance of the target	Disagree(TalkType, day, id) I disagree with utterance of the target

A. Action decisions of one agent model

The one agent model only makes use of the one's self model; in this model, the agent acts based on his or her own will.

1) Estimating roles:

In the one agent model, the agent estimates the role of another player by using the one's self model. We assume that a seer is the player that has the highest probability of a seer, as obtained from the one's self model. Much the same is true of the medium. In this subsection, we describe the player with the highest probability of being a seer as the "assumption-seer"; likewise we describe the player with the highest probability of being a medium as the "assumption-medium." Both the assumption-seer and the assumption-medium are updated before their own utterances and actions, and if they are renewed, the agent provides an utterance by using the estimate given in Table III.

2) Voting:

If the agent belongs on the villager side, aside from the assumption-seer and assumption-medium, the agent votes for the player (from among living players) with the highest probability of being a werewolf in the psychological table.

If the agent's role is possessed, the agent votes for the player with the highest probability of being a werewolf in the psychological table of the role that the agent is imitating; however, the agent avoids the player with a probability of being a werewolf of 1.00 in the psychological table corresponding to the actual role.

If the agent's role is a werewolf, this is essentially the same as begin possessed player, but the friendly player votes only if the player has been clearly identified as a werewolf.

Given the above, the agent states whether he or she thinks the role of the player is a werewolf by using the estimate function shown in Table III above, and then states that he or she votes for the player by using the vote function, which is also shown in Table III above.

3) Special abilities:

If the agent's role is a seer, the agent selects the player with the highest probability of being a werewolf in the psychological table from among players who have not yet divined or expressed that they are a seer as the target of the divination. If the agent has expressed the behavior of a seer, he or she states results of divination that have not yet been uttered by using the divined function shown in Table III above.

If the agent's role is a medium and he or she has expressed the behavior of a medium, he or she states the results of an inquest by using the inquested function shown in Table III above.

If the agent's role is a bodyguard, if the assumption-seer is alive, the bodyguard guards the assumption-seer.. If the assumption-seer is dead and the assumption-medium is alive, the bodyguard guards the assumption-medium. If both the assumption-seer and the assumption-medium are dead, the bodyguard guards the player that has been divined to be a human from among other players, but not player targeted by the given vote. Otherwise, the agent guards a player randomly selected from among other players, but again not the player targeted by the given vote. A bodyguard does not express his or her role, and does not state the target being guarded.

4) Fake special abilities: If the agent's role is a possessed, based on results of a fake divination, the targeted player is randomly selected from among living players, and if the player has expressed being a seer or has utterances inconsistent with his or her own results of the fake divination, results of the fake divination indicate that the player is a werewolf. Otherwise, results indicate that the player is a human.

Regarding results of a fake inquest, if the player that was excused had expressed being a medium or had utterances inconsistent with his or her own results of the fake inquest, results of the fake inquest indicate that the player is a werewolf. Otherwise, results indicate that player is a human.

If the agent's role is a werewolf, regarding results of a fake divination, the player targeted by the divination is randomly selected from among living players. Here, the agent selects three players with higher probabilities of being a werewolf from the psychological table which supposed that one' self is a seer. If the player targeted by the divination is included in these three players, results of the fake divination indicate the player is a werewolf. Otherwise, results of the fake divination indicate the player is a human.

The same logic applies to results of a fake inquest.

Finally, utterances regarding fake special abilities have the same logic as described in Section V-A3.

5) Target of attack:

The agent preferentially attacks the player with the highest sum of the probabilities of being on the villagers' side; however, for players that express the role, the attack is directed to players one by one each in seer and medium.

B. Action decisions of the multiple agent model

Multiple Model Agent have both oneself-model and others-model, and decide a action by inferring the intention of the others. The multiple agent model incorporates both the one's self model and the others model, determining an action by inferring the intention of others

1) Estimating roles:

In the multiple agent model, the agent estimates the role of others by using the one's self model and the others model. We

assume that a seer is the player that has the highest probability of a seer, as obtained from the one's self model and the others model. Much the same is true of the medium. We explain the selection of the player that thinks the most like a seer as an example. First, the agent refers to the probability of the player that expresses being a seer from the psychological table of living players, but not players that express being a seer in the others-model, and calculating the product of this probability and the confidence rating of the psychological table referred to in the others model. The agent performs this calculation for all psychological tables in the others model of all living players, calculating the average value, which is the seer-score of the player that expresses being a seer. The agent calculates this seer-score for all players that express being a seer, considering the player with the highest seer-score as the seer. The agent selects the most reliable medium using the same calculations with respect to a medium. As noted above, we describe the player with the highest probability of being a seer as the assumption-seer, and describe the player with the highest probability of being a medium as the assumption-medium. Similarly, the agent selects an assumption-werewolf from among the players that are living, but not including the assumption-seer or assumption-medium. We use the same approach as that described in Section V-A1 for updates and utterances of the assumption-seer and assumption-medium.

2) Voting: The agent selects the target of a vote by using the werewolf pattern from the others model. If three or more players have expressed this role, a player other than assumption-seer and assumption-medium is considered to be a possessed or werewolf player. A possessed player is more likely to express a fake role, because the death of a possessed player does not directly cause a loss in the werewolf camp. Therefore, we use the method described in Section V-B1 to select the assumption-possessed player (i.e., the player considered to be most possessed) from among players that have expressed a role , but this does not include the assumption-seer and assumption-medium. From the above, players that have expressed a role (other than the assumption-seer, assumption-medium and assumption-possessed) are more likely to be a werewolf. Therefore, the agent selects the werewolf pattern with the highest confidence rating from among werewolf patterns, including the player that is more likely to be a werewolf in the others model. If the player that has expressed a role is less than three, the agent selects the werewolf pattern with the highest confidence rating from among werewolf patterns that include the assumption-werewolf. The agent votes for a player that is living from among players that are included in the selected pattern.

If the agent's role is that of a possessed player, the agent votes for the player for which the sum of probabilities of the werewolf camp is 1.00 from the psychological table of the role that the agent is falsifying. If there is no such player, the agent selects the werewolf pattern with the highest confidence rating from among werewolf patterns that includes the assumption-werewolf. The agent votes for the player with the highest probability of being a werewolf from among living players, but this does not include the werewolf pattern. The player with the highest probability of being a werewolf is selected using the same method as that described in Section V-B1.

If the agent's role is a werewolf, the agent votes for the player that is inconsistent with the agent. If there is no such player, the agent selects three players for which the probability of being a werewolf is highest. The agent then votes for the player from among these three players for which the average probability of being in the villager camp is the highest in others model.

3) Special abilities:

If the agent's role is a seer, the agent selects the divination target by using the werewolf pattern described in Section V-B2. If there is a player that is not yet divined living in the werewolf pattern, the player is divined. If there is no such player, if the assumption-werewolf selected in Section V-B1 has not yet divined, the player is divined. Otherwise, the agent divines the player by selecting at random from players that have not yet been divined. This is the same as that described in Section V-A3 for utterances of the results of divination.

If the agent's role is a medium, the agent uses the same approach as described in Section V-A3.

If the agent's role is a bodyguard, the agent uses the same approach as described in Section V-A3 for selecting the target to guard and the utterances to use. Not that the assumption-seer and assumption-medium are selected using the technique described in Section V-B1

4) Fake special abilities:

If the agent's role is a possessed player, the divination target is the player with the highest probability of being a werewolf, as calculated using the method described in Section V-B1 from among the players, but not the players is included the werewolf pattern that is used in Section V-B2 . If the selected player has expressed being a seer or medium and has utterance that are inconsistent with the results of the agent's divination that player is identified as a werewolf. Otherwise, that player is identified as a human.

Regarding the results of a fake inquest, this is the same as the one agent model, because the inquest target is the player executed the day before.

If the agent's role is a werewolf, regarding the utterances resulting from a fake divination, first, the agent selects the target of the fake divination from among the players that are living and not yet divined. If we assume that the agent utters the results of divination of each player, the agent chooses the player in which the average probability of being in the villager camp from the others model is the highest. The agent then utters the divination results of a selected player. Regarding these divination result, the agent identifies three players with higher probabilities of being a werewolf from the psychological table which supposed one's self that to be a seer, and identify the other players as humans.

Regarding the results of a fake inquest, the agent uses the same approach as that of the one agent model.

Furthermore, for the utterances resulting from divination, the agent uses the same approach as described in Section.V-A4

TABLE IV
WINNING PERCENTAGES FOR EACH ROLE

Player	Vill	Seer	Med	Body	Poss	Wolf
OneModel	0.493	0.472	0.468	0.519	0.531	0.503
MultiModel	0.496	0.518	0.493	0.527	0.554	0.517

5) Target of attack:

The target of attack is the same as that of the one agent model, because the players other than the werewolf obtain information only of the player who is attacked i.e. the players other than the werewolf do not know who attacked, as per the rules of the Werewolf Game, and there is almost no change here other than the player that has been attacked based on the others model.

VI. EXPERIMENTATION

A. Experimental overview

We performed comparative experiments involving the one agent model and the multiple agent model to show the usefulness of the others model. Here, we used the AIwolf server (version 0.2.5) [2]. Opponent Agents are agents that used in the finals of the AIwolf competition in CEDEC2015. We experimented by replacing the agent created for the tournament with the agent was created in this study. We assigned one point to players who have been included in the winning camp, regardless of the life and death, and set up a winning percentage based on the percentage of the total number of trials. We executed 10,000 games in which each role was predefined.

B. Experimental outcomes

From our experiments, we obtained the winning percentages shown in Table IV when executing 10,000 games with each role being predefined.

From this results, when each role was predefined, the winning percentages of the multiple agent model were higher in all roles as opposed to that of the one agent model. In particular, we note that a seer and a medium had significant changes as compared to the other roles.

VII. CONSIDERATIONS

We calculated the rate at which each agent voted for a werewolf and the rate at which each seer divined a werewolf, graphing the rates at which each agent voted for a werewolf at the time of the villager, seer, medium, and bodyguard roles. We show these graphs for the villager, seer, medium, and bodyguard in Figs.2, 3, 4, and 5, respectively. We also show the graph of the rate a seer divined a werewolf in Fig. 6.

These graphs contain a horizontal axis that represents days and a vertical axis that represents rates.

For the villager although no significant differences were found in the winning percentage shown in Table IV, Fig.2 suggests that the percentage of which villager voted for a werewolf in the multiple agent model was larger than that of the one agent model as the game progressed. From these

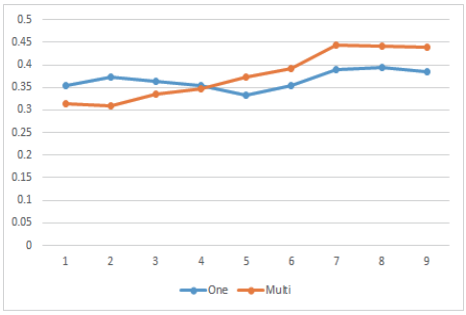

Fig. 2. Percentage for which a villager successfully voted for a werewolf in the one agent and multiple agent models

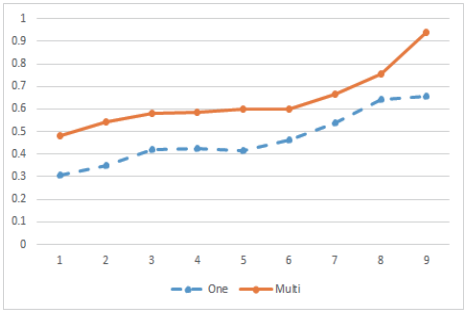

Fig. 3. Percentage for which a seer successfully voted for a werewolf in the one agent and multiple agent models

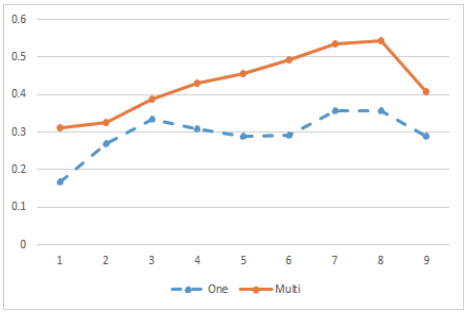

Fig. 4. Percentage for which a medium successfully voted for a werewolf in the one agent and multiple agent models

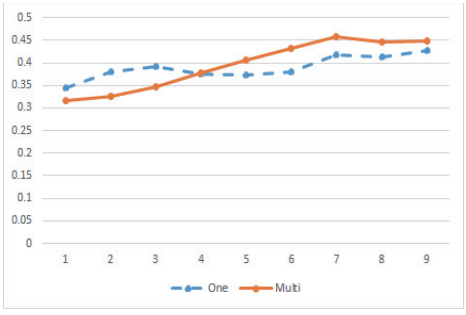

Fig. 5. Percentage for which a bodyguard successfully voted for a werewolf in the one agent and multiple agent models

results, we observe that as the game progressed and the number

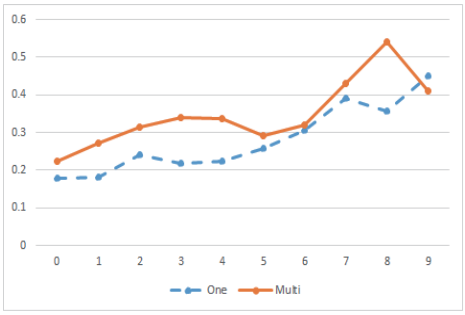

Fig. 6. Percentage of a seer has divined a werewolf

of players decreased, the amount of information increased. Moreover, we note that the multiple agent model was able to find more werewolves.

For the bodyguard, we do not see a big difference in winning percentages in Table IV. Observing Fig.5, we note that the percentage for which a bodyguard voted for a werewolf in the multiple agent model was larger than that of the one agent model as the game progressed. These results are the same as the villager, because although the bodyguard is able to guard one player from attack by a werewolf, it cannot directly know the role of others as a seer or medium. Therefore, we consider this why the results are rather similar to that of the villager.

For the seer and medium, we observe in Table IV that a large difference in the winning percentage has occurred between the one agent model and the multiple agent model as compared to other roles. Inspecting Figs.3 and 4, we observe that the multiple agent model is able to vote for many more werewolves than that of the one agent model. Furthermore, reviewing Fig.6, the multiple agent model is able to divine many more werewolves than that of the one agent model. We conclude here that these have a significant effect on the winning percentage probability. The seer and medium can more efficiently be deleted from the psychological table of the werewolf pattern in the others model, because they can know whether a certain player is a werewolf or not by using their special abilities each day of the game. Therefore, we consider that they were able to more accurately estimate the werewolf.

VIII. Conclusion

In this study, to realize an AIwolf agent that imitate the play of humans, we constructed not only a one's self model that models the role of others as viewed from their own point of view, but also an others model that models the role of others as viewed from others' point of view. We updated the psychological model by using expressions of the role, results of special abilities, votes, utterances, and decided actions by using this model. We performed a comparative experiment of the one agent model that consists only of the one's self model and the multiple agent model that incorporates both the one's self model and the others model. From our results, we found the winning percentage of the multiple agent model to be higher than that of the one agent model for all roles. This occurs because the multiple agent model can estimate the thinking of others based on the action of others. In our experimentation, we showed the usefulness of the others model by focusing attention on strength; however, to realize an AIwolf agent that can play the Werewolf Game while naturally communicating with a human, it remains necessary to pursue how to further model humanity. Therefore, we consider it necessary to experiment with and understand how to better model humanity as a future study.

Acknowledgment

This study received a grant of JSPS Grants-in-aid for Scientific Research 15K12180.

References

[1] H. Osawa, F. Toriumi, D. Katagami, K. Shinoda, and M. Inaba, "Designing protocol of werewolf game : Protocol for inference and persuasion," *FAN Symposium : Intelligent System Symposium-fuzzy, AI, neural network applications technologies*, vol. 2014, no. 24, pp. 78–81, 2014.

[2] F. Toriumi, K. Kajiwara, H. Osawa, M. Inaba, D. Katagami, and K. Shinoda, "Development of ai wolf server," in *Proceedings of the Game Programming Workshop 2014*, 2014, pp. 127–132.

[3] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*. MIT press/Bradford Books, 1995.

[4] T. Bugnyar and K. Kotrschal, "Observational learning and the raiding of food caches in ravens, corvus corax: is it 'tactical' deception?" *Animal Behaviour*, vol. 64, no. 2, pp. 185–195, 2002.

[5] D. Premack and A. J. Premack, *Original Intelligence: Unlocking the Mystery of Who We Are*. McGraw-Hill, 2003.

[6] K. McCabe, D. Houser, L. Ryan, V. Smith, and T. Trouard, "A functional imaging study of cooperation in two-person reciprocal exchange," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11 832–11 835, 2001.

[7] C. Zimmer, "How the mind reads other minds." *Science*, vol. 300, no. 5622, pp. 1079–1080, 2003.

[8] A. Yokoyama and T. Omori, "Modeling of human intention estimation process in social interaction scene," in *2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2010, pp. 1–6.

[9] M. Takano, M. Kato, and T. Arita, "A constructive approach to the evolution of the recursion level in a theory of mind," *Cognitive Studies*, vol. 12, no. 3, pp. 221–233, 2005.

[10] Y. Yasumura, K. Oguchi, and K. Nitta, "Negotiation strategy of agents in the monopoly game," in *International Symposium on Computational Intelligence in Robotics and Automation*. IEEE, 2001, pp. 277–281.

[11] M. Guhe and A. Lascarides, "The effectiveness of persuasion in the settlers of catan," in *IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2014, pp. 1–8.

[12] D. P. Taylor, "Investigating approaches to ai for trust-based, multi-agent board games with imperfect information; with don eskridge's"the resistance"," *Discovery, Invention & Application*, no. 1, 2014.

[13] M. Braverman, O. Etesami, and E. Mossel, "Mafia: A theoretical study of players and coalitions in a partial information environment," *The Annals of Applied Probability*, pp. 825–846, 2008.

[14] E. Yao, "A theoretical study of mafia games," *Arxiv preprint arXiv:0804.0071*, 2008.

[15] P. Migdał, "A mathematical model of the mafia game," *Arxiv preprint arXiv:1009.1031*, 2010.

[16] B. Xiaoheng and T. Tetsuro, "Human-side strategies in the werewolf game against the stealth werewolf strategy," 2016.

[17] G. Chittaranjan and H. Hung, "Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 5334–5337.

[18] F. Xia, H. Wang, and J. Huang, "Deception detection via blob motion pattern analysis," *Affective Computing and Intelligent Interaction*, pp. 727–728, 2007.

[19] L. Zhou and Y. Sung, "Cues to deception in online chinese groups," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, 2008, pp. 146–146.