Proceedings of the Fifteenth AAAI Conference on Artificial
Intelligence and Interactive Digital Entertainment (AIIDE-19)

A Study of AI Agent Commitment in One Night Ultimate Werewolf with Human Players

Markus Eger
Universidad de Costa Rica
markus.eger@ucr.ac.cr

Chris Martens
NC State University, Raleigh, NC, USA
crmarten@ncsu.edu

Abstract

Social deduction games are a genre of board games in which a group of players is secretly assigned roles and each player tries to determine the other players' roles. However, some roles have an incentive to not be found, and the games typically allow players to lie freely. Playing such games is a challenging task for AI agents, because they need to not only determine the probability that each statement made by the other players is truthful, but also come up with convincing lies themselves. In this paper, we present AI agents designed to play one particular such game, One Night Ultimate Werewolf, with human players. We discuss the different deliberation strategies our agents use to determine what they should say, and when they should change their plan. To determine how these different deliberation strategies are perceived by human players, we performed an experiment in which participants played a Unity implementation of the game with each of the three deliberation strategies. We present the results of this experiment, which show that commitment to plans has a measurable effect on player perception and provide a trade-off between consistency and potential for high performance of the agent.

Introduction

Games which feature communication with human players provide a unique challenge for game AI researchers, because the idiosyncrasies of human communication need to be taken into account in order to play the games well. One genre of games in which communication plays a central role is that of social deduction games, such as Mafia/Werewolf[1]. In these games, every player is secretly assigned to one of two factions (Mafia and Citizens in Mafia, Werewolves and Villagers in Werewolf), where one group has an information advantage, and the other has a numerical advantage. Typically, about one third of the players are assigned to the Mafia/Werewolf faction, and while roles are assigned secretly, the Mafia/Werewolf players are told which other players are on the same faction as them. The goal for the Citizen/Villager faction is to deduce who is on the Mafia/Werewolf faction, and vote them out, while the Mafia/Werewolf players try to stay undetected. To facilitate this, there are certain bonus roles, such as a Seer character on the Citizen/Villager team which can learn which team other players are on. The key mechanic, however, is free-form communication, in which players can accuse other players of being on a certain side, exchange information they may have, such as what they have learned using their special roles, or sow confusion, since no player is bound to tell the truth at any point.

Because of the elements of knowledge gathering and exchange, as well as lies and deception, social deduction games provide an interesting application for game AI. While there are many aspects that go into the game, including how human players behave, and which non-verbal cues they exhibit, when they are attempting to deceive other players, our focus lies on the information exchange part of the game, and how it can be modeled by an AI agent. There are several games in the Mafia/Werewolf family, of which we have chosen to work with One Night Ultimate Werewolf. This variant differs from the main game in that there are several roles that not only can gather information, but also manipulate the faction-affiliation of other players, usually without their knowledge. This gives rise to several interesting reasoning phenomena. Over the course of the game, the agent must constantly reevaluate which faction they are on, and if it serves their interests to stick to one story, or change their behavior in the light of new, potentially faulty, information. Additionally, other players may question their statements, which may require the agent to reinforce or explain their behavior. At the end of the game, each player gets one vote to determine who to vote out, for which the agent has to determine which other player is most likely to be on the opposing faction.

In this paper, we present an AI agent design for playing One Night Ultimate Werewolf with a human player. The agents utilize Dynamic Epistemic Logic to reason about the communicative actions of the player, is able to answer the player's questions if it fits with their own plan, and will attempt to deduce their own and the other players' faction af-

[1]Historically, Mafia was invented by Dmitry Davidoff in 1986, and several variants of the game were developed over the years, including Werewolf in the 1990s, which simply replaces the theme but is otherwise identical to the original game

filiation and vote accordingly. Our agents perform reasoning about communicative goals they have, and how long they should be committed to their goal, depending on what they learn from the other players. To determine the effect different levels of commitment on the perception of the agents by human players, we performed an online experiment where participants were asked to play a Unity implementation of One Night Ultimate Werewolf with three different agent types. We provide a summary and interpretation of the most interesting results, and their implications for future work.

One Night Ultimate Werewolf

One Night Ultimate Werewolf (Alspach and Okui 2014) is a Social Deduction game for 5 to 10 players, in which some players are assigned to a Werewolf faction and others to a Village faction. At the beginning of the game, the role cards are shuffled and one dealt to each player, with three extra cards dealt to the center of the table, as shown in Figure 1. Some players on the Village faction may have a role with special abilities that allows them to look at and/or exchange other players' role cards, or interact with the center cards, while others are Villagers without any special ability. When the cards have been dealt, all players may look at their own card once, and game play then consists of three phases:

- A night phase, in which all roles with special abilities perform their special action in a predefined order. In the tabletop version of the game, this is done by all players closing their eyes, and a game master (or smartphone app) calling the roles in a predefined order. When a role is called, the player or players with that role open their eyes (referred to as "waking up") and perform their special action, and then close their eyes again. While a player's role cards may change over the course of the night, they always perform the action associated with the role they started with.
- A day phase, in which players discuss their own faction affiliation, information or suspicions they may have about other player's faction affiliation, or inquire about any information they believe another player might have. For any of these exchanges, players are free to lie partially or fully. In the tabletop version, the day phase is limited to a certain time, which is typically around 5 minutes.
- A voting phase, in which all players simultaneously vote for another player. If any player gets more than one vote, the player or players with the most votes are voted out. If at least one Werewolf is voted out this way, the Village faction wins the game, otherwise the Werewolf faction wins. Players win or lose with the faction that corresponds to the role card that they have at the end of the game, which may be different than the one they started with.

The game, and its various expansions, provide about 80 different role cards, of which we will only describe the ones supported in our implementation of the game:

- The **Werewolves**, which should be about one third of the players, all wake up together, so each of the Werewolves knows who the other Werewolves are.

Figure 1: Game setup for One Night Ultimate Werewolf. For an actual game, the role cards for each player, as well as the center cards, are dealt face-down.

- The **Seer** wakes up and can look at another player's card, or two of the center cards.
- The **Rascal** wakes up, and may exchange their two neighbors' cards.
- The **Robber** wakes up and may take any other player's card and exchange it with their own, and look at their new card.
- The **Insomniac** wakes up and looks at their own card.

The challenge the game poses to the players is that they constantly have to question their own faction affiliation as well as the statements of the other players. For example, if Anna starts as the Werewolf, Brian, who is the Rascal, sits to their right, and Carol, who is the Insomniac, is on other side of the Rascal, the Rascal may exchange the Werewolf player's card with the Insomniac's card. This exchange would cause Anna to be on the Villager team, and Carol to be on the Werewolf team. During the day phase, when Brian claims to be the Rascal and to have swapped his two neighbors' cards, Anna needs to consider whether to believe Brian or not. If they believe Brian, they might state that they started as the Werewolf, which, in turn, the other players have to evaluate for its believability. However, if Dave started as the Seer, he may have additional information, perhaps because he looked at Anna's card. Generally speaking, the game works because the Werewolves constitute a minority among the players, which means the Villager faction can collectively determine who the Werewolves are, if they manage to corroborate each other's stories. We will now present some approaches to AI agents for social deduction games.

Related Work

Social deduction games provide many interesting challenges for AI agents, and several researchers have looked into different aspects. While our work was on One Night Ultimate Werewolf, most work on its ancestors Mafia and Werewolf is just as applicable. For example, Chittaranjan and Hung (2010) have investigated how to use pitch and tone of voice to determine whether players are lying about their roles. For our work we eschew direct interaction with verbal commu-

nication in favor of fixed statements, in order to be able to focus on the reasoning about beliefs caused by the content of the communicative actions. Gillespie et al. (2016) have shown that the statements made by players typically fall into a set of semantic categories, and Hancock et al. (2017) describe how these statements are formed into plans that may or may not have deception as a goal. In contrast, our work utilizes a measure of plan quality to move beyond a binary notion of deception towards a measure of how much a plan may convince other people of a statement. Other researchers have looked into developing actual strategies for Werewolf, such as Braverman et al. (2008), who use a probabilistic model to determine which setups lead to fair games. However, they utilize a simplified version of the game, which actually removes most communication from the game, in order to be able to derive theoretically optimal strategies. Bi and Tanaka (2016) describe how one of these strategies can be exploited by the villagers, if the Werewolf players are assumed to behave exactly like Villagers without any special role. Finally, Nakamura et al. (2016) describe an approach to Werewolf in which an AI agent forms a model of the beliefs of the other players in order to deduce probabilities for their roles. However, their model requires expert players to provide estimates for probabilities for different combinations of roles and actions.

Our work is mainly based on a formulation of intentionality by Cohen and Levesque (1990), based on prior work by Bratman (1990). They describe intention as choice with commitment, which means that agents make decisions about which goals to pursue, form plans to achieve them, and then reevaluate how long to follow each plan, and when to adopt a new goal. However, the problem of when to maintain and when to drop intentions is non-trivial (Rao and Georgeff 1995), and therefore several different approaches have been proposed. As Braubach et al. (2004) discuss, one challenge is that goals are often talked about in abstract terms, leaving a gap between the theory of adopting, pursuing and dropping them, and the actual implementation. To address this, their system uses a classification of goals into different categories and a formalism to describe different goals, but they also leave the actual deliberation strategy open for future work. Pokahr et al. (2005) build on this work and present what they call the *Easy Deliberation Strategy* that uses the information contained within goals to choose between them, without regard for the actual plans. Mohanty et al. (1997) describe a different approach, in which an agent is influenced into adopting plans by other agents, which is more useful in social settings, such as an employee taking orders from their supervisor. Perhaps the most concrete agent design, that combines goal selection with a reconsideration-mechanism was described by Wooldridge (2000). His agents have a notion of beliefs, which they use to determine which intentions to adopt, but over the course of plan execution the agent's beliefs may change, which causes them to reconsider their intentions. Our work expands upon this by incorporating not just agent beliefs, but also beliefs about the beliefs of other agents using Baltag's variant of Dynamic Epistemic Logic (Baltag 2002), using our implementation Ostari (Eger and Martens 2017). Ostari is implemented in Haskell, and allows the specification of actions that affect agent beliefs, and also provides the capabilities to plan to achieve goals involving such beliefs. We have previously implemented One Night Ultimate Werewolf in Ostari (Eger and Martens 2018), and the present work is a direct extension of this effort, as described in the next section.

Approach

Our approach is an extension of previous work we presented which focused on comparing AI agents for One Night Ultimate Werewolf that played simulated games with each other to determine the effect of different levels of commitments to plans (Eger and Martens 2018). We will briefly summarize how these agents operated, to be able to discuss the changes we made in order to enable the agents to play with a human player. We will also briefly present the architecture that constitutes the connection between our Unity front-end and the Dynamic Epistemic Reasoning system Ostari (Eger and Martens 2017) on the back-end.

Note that our implementation of One Night Ultimate Werewolf made two key changes from the tabletop version: First, the discussion phase is performed in a turn-based fashion, with a fixed limit of 7 turns, where each player can perform up to one communicative action per turn, instead of the time limit present in the tabletop version. Second, while the tabletop version allows players to make arbitrary statements, our implementation provides the players with a fixed set of statements, that allows them to discuss any aspect of the game state. We made these changes since natural language processing was beyond the scope of our work, and to be able to focus on the planning aspects of the game.

Existing Agent Design

To play One Night Ultimate Werewolf, we built agents that utilize Dynamic Epistemic Logic (Van Ditmarsch, van Der Hoek, and Kooi 2007) to represent communicative actions, and then reason about how to best reach a goal using these actions. In Dynamic Epistemic Logic, beliefs are represented as possible worlds, meaning that each combination of facts an agent considers possible constitutes one world. In a game like One Night Ultimate Werewolf, shuffling the cards creates one possible world per card permutation for each agent, because the agents are unaware of the order of the cards. When an agent looks at a card, they eliminate all worlds that are inconsistent with the card they saw. On the other hand, actions that happen without an agent being aware of its exact parameters, such as another player possibly exchanging two cards, cause the agent to add possible worlds for each variation of the action the agent considers possible. However, communicative actions do neither remove worlds, since they do not constitute certain knowledge about the validity of any particular worlds, nor do they add worlds, since the agent must already consider what is said possible beforehand to believe the speaker. Instead, our agents use the observation that the Village team has an interest in telling the truth, and since they constitute the majority of the players the majority of statements in a typical game should be truthful. Whenever a statement is made, our agent marks

all worlds they consider possible that would be inconsistent with that statement as being less likely. The aggregate effect over several statements is that worlds that are consistent with more statements are considered more likely than worlds that would contradict more statements made by the players. We call this measure of how likely worlds are the *weight* of each world, and use it to calculate a *weighted quality* of beliefs that depends on how many worlds a statement holds in, and which weights these worlds have, compared to the sum of weights of all worlds.

The agents use this weighted quality of beliefs to form plans involving communicative actions, that they believe to change the weighted quality of beliefs of other agents according to their goals. For example, a player that believes that they are on the Werewolf faction may have a goal to convince other players that they are a Villager. Our agents represent this goal as trying to change the weighted quality of another player's belief that the agent is a Villager above a threshold value. The quality of a plan consisting of a sequence of actions can then be defined as how well it achieves that goal, i.e. what the resulting weighted quality of the other player's belief will be after the plan is executed. Our agents use a list of expert-provided candidate goals, find the highest quality plan to achieve each goal, and then adopt the goal which can be achieved with the highest quality. On subsequent turns, the agent will reevaluate their decision, by computing new plans for each goal, and then switch to that new plan depending on the deliberation strategy. For this project, we used three different deliberation strategies, called capricious, balanced and fanatical. A capricious agent will disregard their existing plan's quality and always adopt the new plan with the highest quality, while a balanced agent will only adopt a new plan if its current quality is higher than the quality of the existing plan as evaluated when the plan was adopted. A fanatical agent, finally, will never reevaluate plans, and always keep pursuing the plan they initially adopted until it is finished. We will now describe the adaptations we made to this agent design in order to allow a human player to play with the agents.

Because the agent can never be certain to convince another player of anything with certainty, adding more communicative actions to reinforce their statements always lead to a better plan. The fanatical agent will therefore always come up with a plan for the entirety of the game, and never change from it. The capricious agent, on the other hand, will reevaluate their plan every turn, but often come up with exactly the same plan they were already following. In a typical game, though, the capricious agent will change their plan between 2 and 4 times. The balanced agent we used for our experiment falls between these two extremes, and will change their plan between 1 and 2 times on average during a game.

Playing One Night Ultimate Werewolf with Human Players

The biggest change from having agents play One Night Ultimate Werewolf with each other to having them play with a human player has to do with interactivity. While the agents themselves will form plans to convey information to the other players, a human player may want to inquire about details, or is otherwise not satisfied with the information volunteered by the agents. In order to address this, we made two changes to our system. First, in order to be able to actually ask something, we added communicative actions that represent questions. Second, we also modified the agents to take these questions into account when performing their communicative actions. However, while we want the agents to be responsive, we also do not want the human player to be able to explicitly control their behavior. We addressed this concern by integrating question answering with our existing goal deliberation process.

Questions, unlike other communicative actions, do not change the listeners belief state, but instead attempt to change their plan. Question actions in our encoding of the game reflect this by setting a question topic for the player the question was directed at, and other communicative actions are tagged with the question topic they address. While the question topic is maintained per player, its value is actually public, since all communicative actions, including questions, are overheard by all players. When it is an agent's turn, and they are about to deliberate on which new plan (if any) to adopt, they now take additional goals into account: For each existing candidate goal, the agent will also consider an additional candidate goal that has the additional condition of addressing their current question topic. When a plan is formed, it may either be in service of the provided candidate goals, or also include communicative actions that address the agent's question topic.

Consider the case where an agent Anna forms a plan to convince the other players that they are the Villager, but they were just asked which role they believe Brian has. When Anna forms a plan to convince the other players of being a Villager, perhaps by simply stating they they are a Villager, they may also add actions that state that they believe Brian to be a Werewolf. Because this plan would addresses Anna's original goal as well as the goal of addressing the question posed to her, it would always have a higher quality than a plan without the additional communicative action about Brian's role. This means, if we just added conditions to all goals, the agents would always prefer to answer questions. However, we want the question answering to be a natural part of the agent's plan, instead of additional actions they add to the end of another plan.

Because the game is played in turns, with a limit of 7 turns, and one action per turn, we limited the agents to constructing plans of a maximum length equal to the number of turns remaining. With this plan length restriction, the agents have to decide between plans that address their actual goals better and plans that include answering questions that were posed to them. Since the agents can never actually execute plans that are longer than the remaining turns of the game, this restriction actually results in better fitting plans, and has the added benefit of reducing the planning time of the agents.

System Design

In order to let humans actually play the game, we developed a Unity front-end, shown in Figure 2, that was exported as a WebGL build playable in a webbrowser. Since the actual game logic was implemented in Ostari, running as a Haskell-

Figure 2: A screenshot of the user interface for the human player.

process on our game server, we developed a web-server in python to serve as the bridge between the front-end and the Ostari back-end. When a player starts the game, this server generates the initial role assignment, and sets up the appropriate agents, generating an Ostari input file. It then starts the Ostari process, and connects to it via pipes. When the player performs an action in the Unity WebGL application, a HTTP request is sent to the web server, which generates the necessary input to perform that action in Ostari and passes it via the input pipe. It then interprets the resulting output and sends the result inside the HTTP response. This setup makes the game more enjoyable for the human player, by providing them with a nice graphical user interface, while still allowing us to use the powerful Dynamic Epistemic Logic reasoning capabilities provided by Ostari. We used this setup to run an evaluation of our agents with human players, as we will describe now.

Results

In this section, we will describe how we evaluated the effect of different levels of commitment on how human players perceived and rated our AI agents. First, we will provide a description of the experiment design and setup, and then follow with a discussion of the most important results.

Experiment Design

For our experiment, we compared three different agent types: Capricious agents, which decide which plan to follow every turn afresh, fanatical agents, which follow their initial plan until it is finished, and balanced agents, which use the measure of plan quality to decide when to change plans. In order to compare the agents, we asked participants to play three games, where each game was played with a different agent type, assigned in random order. Each game was played with one human player and 4 AI agents, where each AI agent was assigned the same agent type.

We also limited the starting assignment of role cards to one of three scenarios: One in which the participant started as a Werewolf, but the Werewolf role could be taken away from them, one in which the participant started as the Seer, but could become a Werewolf, and one in which the participant started as the Rascal, with a Werewolf to their left, and the Insomniac to their right, so their decision on whether or not to change their neighbors' cards would affect which player would be the Werewolf. As with the agent type, these three starting configurations were assigned in random order. After each game, each participant was asked a series of survey questions:

- Whether the AI agents changed their behavior based on the player's actions.
- If the AI agents' actions made sense.
- If the AI agents played well.
- If it was fun to play with the AI agents.
- Which of the statements of one of the AI controlled players they believed at the time they were made.
- Which of the statements of one of the AI controlled players they considered to have made sense at the time they were made.

The first four questions were rated on a 5-point Likert scale, while for the last two actions participants were shown the statements in question and asked to select all that applied. Additionally, participants could give free-form text feedback about the agents' behavior after each game. After three games, participants were additionally asked for some basic demographic information, such as age, board game experience in general, and with One Night Ultimate Werewolf in particular, as well as their estimate for which percentage of games is won by the Werewolf faction in a regular game of One Night Ultimate Werewolf.

Experiment Results

Participants for the experiment (NCSU IRB # 14087) were recruited on social media, including boardgamegeek.com and the boardgames subreddit, and the experiment was run online for 2 weeks. In this time, 71 participants finished at least one game and answered the survey questions corresponding to that game. As expected, the population skewed towards participants with an affinity for board games, with 42 self-identifying as gamers, and only 5 not, with the rest choosing to not answer the question. For each of the six survey questions, we performed a Mann-Whitney-Wilcoxon test to test if the response for an AI agent type could be expected to be higher than for the other two, and a χ^2 test to test for a difference in distribution between the three different agent types. A Holm-Bonferroni correction was used to account for multiple testing. We also qualitatively analyzed the free-form comments from the participants.

Interestingly, while the means were not found to be statistically significantly different for any of the survey questions, several questions showed a difference in distribution. The number of statements players rated as making strategic sense differed between the balanced and the fanatical agent ($p < 0.001$, $\chi^2 = 29.23$), as well as between the balanced and the capricious agent ($p = 0.002707$, $\chi^2 = 21.84$). Figure 3 shows a histogram of how many participants rated 0, 1, 2, etc. statements made by each agent type as making sense.

As the statistical tests tell us, the distribution of how many statements are considered to make sense differ between the agent types, with the balanced agent having significantly fewer games in which none of its statements made sense, and several where 5 or more did. On the other hand, the fanatical agent had few games in which 5 or 6 of its statements made sense, but more in which all 7 of its statements were rated as making strategic sense. The capricious agent, on the other hand, has very few games in which all of its statements made strategic sense.

These results are in line with our expectations, since the fanatical agent makes up a plan at the beginning of the discussion phase and then keeps following it to the end. In games in which nothing contradicts the facts conveyed by that plan it will make perfect sense, but this is not always the case. On the other hand, the capricious agent may change their story at any time, making it too erratic, causing the human player to find statements to be inconsistent. The number of statements players said they believed was also statistically significantly different between the fanatical and the balanced agent in distribution ($p < 0.00001$, $\chi^2 = 39.3$), with a similar bimodal distribution for the fanatical agent, where the participants believed either all or none of the statements, and a more balanced distribution for the balanced agent, where there were fewer games in which none of the statements were believed, but also fewer in which all of them were.

Another interesting result was how players rated the skill of the AI agents, as shown in Figure 4. While the χ^2 test revealed a difference in distribution between the fanatical and balanced agent ($p = 0.003416$, $\chi^2 = 15.722$), as well as between the capricious and the balanced agent ($p = 0.006486$, $\chi^2 = 14.27$), these results should not be taken as statistically significant due to the necessary Holm- Bonferroni correction. However, since the Holm-Bonferroni correction does not make any assumptions about the dependence structure of the different tests, it is known to lead to type II errors in cases where the different responses are positively correlated (Abdi 2007). We believe that such a correlation is likely, and that further experiments are needed to get more conclusive results.

Qualitatively, we also looked at the free form text responses the participants provided. Because of the higher effort required to provide such feedback, only between 21 and 26 provided answers for each agent type. Several themes were reoccurring across multiple answers, including a frustration that the AI agents did not respond to questions, which 3 participants noted for the capricious agent, 4 for the balanced agent and 8 for the fanatical agent, with one participant writing *They do not respond to questions well if at all. It would be more fun to play with them if it felt like they could react to what I asked.* after their game with the fanatical agent. Another particularly interesting theme was that the voting behavior of the AI agents confused the participants, with 9 participants noting this for the capricious agent, 7 for the balanced agent, and 1 for the fanatical agent. We believe this is due to the fact that the agents perform their voting purely based on their estimate on the game state, which they do not fully communicate to the player.

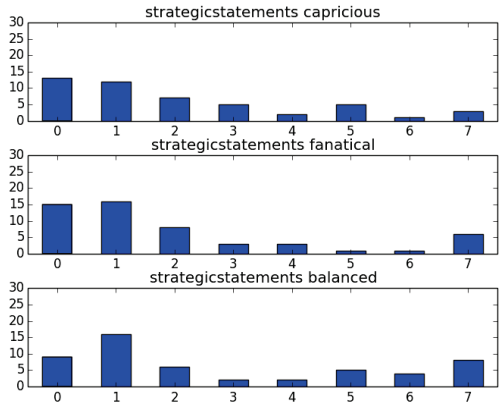

Figure 3: Histogram of the number of statements the participants rated as making strategic sense for each of the agent types.

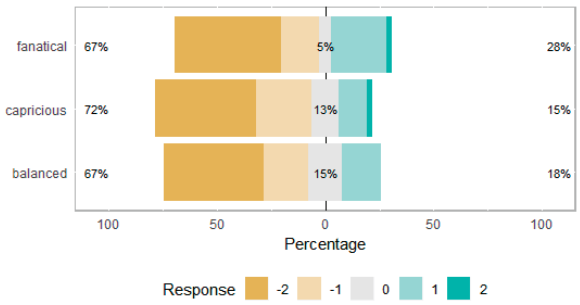

Figure 4: Participant rating for the skill of the AI agents from -2: very bad, to 2: very good.

Conclusion and Future Work

We have presented agents that play the social deduction game One Night Ultimate Werewolf with human players, based on our previous work. The main changes over previous work are enabling of question answering in order to make the agents more interactive for the human players, and providing a graphical user interface implemented in Unity. We used this user interface to perform an experiment to determine how human players evaluate the commitment of agents to their plans. While the participants did not find the agents to behave differently on average, there were interesting differences in the distribution of the ratings. In particular, agents that would commit to a goal fanatically and never change their plans were perceived as either very bad or as very good, while agents that more carefully decided when to change plans had fewer exceptionally bad games, but also fewer exceptionally good ones. This provides additional evidence that agent commitment matters to the perception of human players, but also shows that a high commitment can lead to potentially achieving highly desirable results, at the increased risk of failing badly.

There are several opportunities to improve upon our work.

While our agents provide different options for trade-offs, and use a wide variety of expert-provided goals, the actual strategic trade-offs presented by the game have not been determined yet. We believe our action encoding could be used to analyze the game more formally in the future. Additionally, while most players found our user interface intuitive, it simplifies the game in some ways. Adding the capabilities for audio-processing and free communication could greatly increase how engaging the game is perceived to be, but would require significant effort that was out of scope for our work. One participant also noted that the lack of non-verbal cues constituted a real limitation, since they like to use facial expressions, as well as voice patterns to determine the truthfulness of statements.

Acknowledgements

This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

References

Abdi, H. 2007. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics* 3:103–107.

Alspach, T., and Okui, A. 2014. One night ultimate werewolf. https://beziergames.com/collections/all-uw-titles/products/one-night-ultimate-werewolf.

Baltag, A. 2002. A logic for suspicious players: Epistemic actions and belief–updates in games. *Bulletin of Economic Research* 54(1):1–45.

Bi, X., and Tanaka, T. 2016. Human-side strategies in the werewolf game against the stealth werewolf strategy. In *International Conference on Computers and Games*, 93–102. Springer.

Bratman, M. E. 1990. What is intention. *Intentions in communication* 15–32.

Braubach, L.; Pokahr, A.; Moldt, D.; and Lamersdorf, W. 2004. Goal representation for BDI agent systems. In *ProMAS*, volume 3346, 44–65. Springer.

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

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

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial intelligence* 42(2-3):213–261.

Eger, M., and Martens, C. 2017. Practical specification of belief manipulation in games. Proceedings of the 13th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment.

Eger, M., and Martens, C. 2018. Keeping the story straight: A comparison of commitment strategies for a social deduction game. Proceedings of the 14th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment.

Gillespie, K.; Floyd, M. W.; Molineaux, M.; Vattam, S. S.; and Aha, D. W. 2016. Semantic classification of utterances in a language-driven game. In *Computer Games*. Springer. 116–129.

Hancock, W.; Floyd, M.; Molineaux, M.; and Aha, D. 2017. Towards deception detection in a language-driven game. In *Proceedings of the Thirtieth International Florida AI Research Society Conference*. AAAI.

Mohanty, H.; Patra, M. R.; and Naik, K. S. 1997. Influencing: A strategy for goal adoption in BDI agents. In *Proceedings of the 2nd International Conference on Cognitive Technology (CT97)*, 175. IEEE Computer Society.

Nakamura, N.; Inaba, M.; Takahashi, K.; Toriumi, F.; Osawa, H.; Katagami, D.; and Shinoda, K. 2016. Constructing a human-like agent for the werewolf game using a psychological model based multiple perspectives. In *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016, Athens, Greece, December 6-9, 2016*, 1–8.

Pokahr, A.; Braubach, L.; and Lamersdorf, W. 2005. A goal deliberation strategy for BDI agent systems. *Multiagent System Technologies* 82–93.

Rao, A. S., and Georgeff, M. P. 1995. The semantics of intention maintenance for rational agents. In *Procs. of 14th International Joint Conference on Artificial Intelligence (IJCAI95)*, 704–710.

Van Ditmarsch, H.; van Der Hoek, W.; and Kooi, B. 2007. *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Wooldridge, M. J. 2000. *Reasoning about rational agents*. MIT press.