Proceedings of the Fourteenth Artificial Intelligence and
Interactive Digital Entertainment Conference (AIIDE 2018)

Keeping the Story Straight: A Comparison of Commitment Strategies for a Social Deduction Game

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

Abstract

Social deduction games present a unique challenge for AI agents, because communication plays a central role in most of them, and deception plays a key role in game play. To be successful in such games, players need to come up with convincing stories, but also discern the truth of statements of other players and adapt to the information learned from them. In this paper we present an approach for virtual agents that have to determine how long to stick to their story in the light of information obtained from other players. We apply this approach to a particular social deduction game, One Night Ultimate Werewolf, and demonstrate the effect of different levels of commitment to an agent's story.

Introduction

The genre of social deduction games is typically characterized by games in which a group of players tries to deduce a subset of the group that is assigned to a stereotypically "evil" faction such as spies, criminals or mythical creatures. The games have different mechanics for some of the actions the players can perform, but typically feature free-form communication between players, and a key part of game play consists of this discussion. In one such game, One Night Ultimate Werewolf (Alspach and Okui 2014), players are assigned to one of two factions, the Werewolves and the Village. Players on the Village faction win if they can identify at least one of the Werewolves, whereas the Werewolves win if they all stay undetected. Players are assigned special roles within these factions that allow them to look at other players' role cards or exchange them. The core of the game play involves communicative actions which the players use to share the information they got, or — for the Werewolf players — come up with convincing alibis. Because the allegiance of players with one of the two factions may change over the course of the game, if their role cards are exchanged with another player's, usually without them knowing, it is essential to determine how long to stick to a story, and when to change strategies. On the other hand, for the non-Werewolf players, convincing other players of their own identity may become less important than communicating a suspicion about another player if one arises from the information obtained by other players.

However, the problem of how long to pursue one goal over another is not just relevant in One Night Ultimate Werewolf and other social deduction games. Rather, it is one of the central challenges in designing intentional agents, or agents that act in a goal-driven manner. In fact, one of the seminal papers in this area, written by Cohen and Levesque (1990), describes intentionality as choice with commitment. This means for an agent to behave intentionally, it needs to choose a goal to pursue, plan how to achieve that goal and then follow that plan as long as is appropriate. However, determining just how long it is appropriate to stay committed to an intention is non-trivial. Constantly switching which goal to pursue when the agent finds a better one makes it behave erratically, which is why the notion of commitment exists in the first place. On the other hand, as the agent obtains new information, it may become apparent that continuing to pursue a goal is sub-optimal or even futile.

In this paper we present a novel characterization of commitment to agent goals that utilizes a measure of closeness to a goal to distinguish between different plans to achieve these goals. This measure is based on the agent's mental model about the state of the world and other players' beliefs, making it particularly well suited for domains that feature communication and other actions that manipulate agent beliefs. We used our definition of commitment to develop an approach to playing a formalization of the game One Night Ultimate Werewolf in Dynamic Epistemic Logic, focusing on the logical ramifications of agents' communicative actions. To demonstrate the impact of different levels of commitment to goals, we performed in-depth simulation experiments with several different combinations of agents and game play scenarios that we also present. Since intentionality of agents play a key role in their interaction with humans, we are planning on using the results obtained in our experiments as the basis for building agents that play One Night Ultimate Werewolf with human players in a way that is enjoyable for them, but we believe a similar approach can also be used in any other application requiring communication.

Related Work

Intentionality is such a core expectation of humans observing or interacting with intelligent agents (Dennett 1971), they will describe agents as acting intentionally even when the agents are not. Gallagher et al. (2002), for example, had subjects play rock-paper-scissors with an agent, and the subjects tried to explain the agent's behavior as intentional even when it was in fact playing randomly. However, the brain activity measured in the human subject differed between an opponent following a strategy and one playing randomly, so even if humans could not tell that there was a difference, they picked up on it on a subconscious level. Mateas (2003) has thus argued that for the development of believable agents for games intentionality plays a key role. Eger et al. (2017) describe how agents that use and communicate intentions achieve a better performance in a collaborative game when compared to agents that simply follow a set protocol. In contrast, our agents play competitive games that feature lying, and while their agents used intentions, plans were limited to one step, while the plans our agents generate have multiple steps, leading to the need to revise them.

As a formal basis for intentionality serve work by Bratman (1990), and Cohen and Levesque (1990). Together, they enumerate several properties of intentional behavior, based on the idea that for an agent to adopt an intention they need to form a plan to achieve that intention and believe that they are actually able to and going to achieve that goal. While they describe how intentions that an agent no longer considers possible are dropped they also acknowledge that agents may reconsider and drop intentions for other reasons as well, without going into detail on how to determine when to do so. They also don't address how an agent decides on which intention to adopt in the first place. Pokahr et al. (2005) describe an approach to this problem by using information solely contained in the goals without regard for the plans needed to achieve them, while Mohanty et al. (1997) use an operationalization of social norms to adopt goals depending on communicative actions by other agents, like in the case of a subordinate following orders from their superior. Wooldridge (2000) describes an entire agent architecture that includes decisions on when to adopt, drop and revise intentions, which serves as basis for our own agent design. However, none of these approaches accounts for goals that can be approximated, as in our work, and assume that goals can only either be reached entirely or not at all.

The game we are using as our application domain, One Night Ultimate Werewolf, is an evolution of a previous game called Mafia by Dmitry Davidoff in 1986, that was rethemed as Werewolf by Andrew Plotkin in 1997[1]. The games are very similar, with the main differences being that Mafia/Werewolf did not involve characters that could change other players' roles, and proceeded over several rounds, eliminating players over time. This makes attribution of which actions have an effect on the outcome significantly harder, especially since the communicative actions and non-communicative actions alternate. It also leads to many degenerate games where one key character is killed arbitrarily early in the game because of a lack of information at that stage of the game. One Night Ultimate Werewolf is much more recent, and a large body of prior work has been performed to investigate different aspects of the original Werewolf game. Because of the similarity of the games, though, much of this work is still applicable to the newer game. Chittaranjan and Hung (2010), for example, use features of the players' speech, such as pitch or speaking rate to detect when the players are lying. Gillespie et al. (2016) describe how utterances in such games can be classified depending on their intended semantics. While our research eschews processing speech to focus on actual game play properties, this work served as the basis for the statement types that are allowed in our implementation.

Because of the complexity of the communicative aspect, research into actual strategies has been limited so far. Braverman et al. (2008) use a simplified version of the game to determine a fair composition of different characters, but this work removes most decisions about communication from the players, and uses a fixed protocol instead. Bi and Tanaka (2016) improve on this work by attacking the particulars of the Werewolf strategy, but this is limited to the case in which the Werewolf follows Braverman's protocol of pretending to be an ordinary Villager. Finally, Nakamura et al. (2016) describe an agent that uses a model of other players' beliefs to determine the probabilities for role-assignments to players based on an expert-provided baseline.

Dynamic Epistemic Logic

When describing the properties of intentionality, Cohen and Levesque often refer to the agent's beliefs. Additionally, it has been shown that having a model of *other* agents' beliefs also provides an advantage in competitive games (De Weerd, Verbrugge, and Verheij 2013). While many authors have proposed models of agent beliefs, in particular for believable characters in narratives (Shirvani, Ware, and Farrell 2017; Teutenberg and Porteous 2015; Thorne and Young 2017), we opted to base our work on a logical framework with formal reasoning semantics. Concretely, Dynamic Epistemic Logic (Van Ditmarsch, van Der Hoek, and Kooi 2007) enables reasoning about player beliefs, including elimination of contradictory worlds. Of the several different flavors of this logic that exist, our work uses one developed by Baltag (2002) because it provides not only a model of beliefs about states, but also of appearances of actions, so that players can perform actions in secret, or suspect that other players perform particular actions. In this model, an *epistemic state* $\mathcal{M}$ is a collection of *worlds* w, each of which represents a particular set of facts being true, with one world w_0 designated as the *actual world* [2]. For each agent a, each world w is associated with an *appearance* $\mathrm{app}_a(w)$ which is a set of worlds which an agent considers plausible alternatives to w. For example, if player b is the Werewolf, but player a does not know that, player a will consider worlds possible in which player b is the Werewolf, worlds in which player c is the Werewolf, etc. An agent a is said to *believe* some proposition ϕ in $\mathcal{M}_0$, written $\mathcal{M}_0 \models \Box_a \phi$ iff ϕ holds in all worlds a considers possible in the actual world, $\mathrm{app}_a(\mathcal{M}_0)$.

[1] http://www.eblong.com/zarf/werewolf.html

[2] Notationally, a state $\mathcal{M}$ is usually identified with its actual world w_0, written $\mathcal{M}_0$

One Night Ultimate Werewolf

One Night Ultimate Werewolf (Alspach and Okui 2014) is a social deduction game. Players are secretly assigned different roles, with $n + 3$ cards dealt to n players, with the remaining cards being placed in the center of the table, some of which belong to the Werewolf faction, while others belong to the Millers Hollow faction [3]. Game play is driven by the goal of deducing the roles of other players. Some roles provide the players with additional information, such as the Seer that can look at the role card of another player. Game play begins with a night phase, during which the different roles can use their special abilities, followed by a day phase which consists of free form discussion between the players. After a set time limit the discussion wraps up, and each player votes for one other player. If a Werewolf gets the plurality of votes this way, the Millers Hollow faction wins, otherwise the Werewolf faction wins. The available roles include:

- The **Werewolves**, which are told which other players belong to the Werewolf faction
- The **Seer**, which can look at any player's role card
- The **Rascal**, which can exchange the two role cards of the two players on either side of the Rascal
- The **Robber**, which can exchange their own role card for any other player's role card, and can look at their new role card
- The **Insomniac**, which can look at their own card at the end of the night phase
- The **Villager**, which have no special ability

One aspect of the game that requires players to reconsider their plans as game play progresses is that some roles, such as the Rascal and the Robber, can exchange players' role cards with those of other players. While players always perform actions during the night phase according to the role they were initially given, they win or lose with the faction of the role card they have at the end of the game. For example, if a player is initially given a Werewolf card, but the Robber steals that card and exchanges it for their own Robber card, the former Werewolf now wins iff the Millers Hollow faction wins, while the former Robber now wins iff the Werewolves win.

A Formal Characterization of the Game's Communicative Actions

The game, as originally designed, provides many interesting challenges for researchers, such as synthesizing and recognizing verbal and non-verbal communication, and proper communication etiquette such as not unnecessarily interrupting other speakers, and determining when to speak. However, for the purpose of this paper we want to focus on the underlying reasoning process the players have to go through to determine the roles of other players and when and how to lie. Therefore, instead of allowing a free-form discussion during the day phase, our variant of the game is turn-based, with a turn limit rather than a time limit and restricts the players to several predefined statements:

- Claim to have started with a certain role card
- Claim to currently have a certain role card
- Claim that another player (or a center card) started with a certain role card
- Claim that another player (or a center card) currently has a certain role card
- Claim to have performed any of the actions available during the night phase
- Don't say anything

These actions were chosen for the players to be able to talk about the entire game state and its history from the start of the game on. This means, if every player tells the truth and believes everything every other player says, the players could all come to know the entire game state using these communicative actions. We implemented the (secret) night time actions, as well as these communicative actions, in Ostari, a macro language that simplified writing epistemic actions (Eger and Martens 2017).

Agent Design

Our agent is based on a design by Wooldridge (2000), adapted to our turn-based domain. The agent keeps track of its beliefs about the world, including beliefs about other agents' beliefs and updates them according to actions performed on other agents' turns. On its own turn, our agent always computes a new plan to pursue the best suitable goal in the current situation and compares it to the plan it is already following (if one exists). The new plan is then adopted as the current plan iff it is significantly better than the current plan, where the exact definition of "significantly better" defines the level of commitment of the agent.

Belief Quality and Lying

Our agents use Dynamic Epistemic Logic as described by Baltag (2002) to model other players' beliefs. At the beginning of the game, each agent is dealt their own role card and looks at it. Then each player considers $(n - 1)!$ worlds possible, where n is the number of cards that were dealt, because they know their own card, but they don't know where each other card ended up. They also know that in each of these worlds each player knows their own card, but not the cards of the other players. Now consider the action of the Seer looking at another player's card: No agent, except for the Seer, knows which card the Seer looks at, but for each possible choice, they know that the Seer now knows the role of that player.

Note that the worlds each agent considers possible have a direct relation to probabilities of something being true. If there are 8 cards, for example, for player a the fraction of worlds they consider possible in which player b has a particular card represents the probability of player b actually having that card. This means, in addition to being able to

[3] To avoid confusion between the Villager role and the Village as a faction, we will refer to the faction by the name of the village featured in a popular Werewolf release, Millers Hollow

determine what an agent actually believes, we can also determine how close they are to believing something. Because this only constitutes a true probability if we assume other players' actions are chosen uniformly at random, we will refer to this measure as the *quality* of a belief[4].

Formally, we define the quality Q of a statement in an epistemic state $\mathcal{M}$ as:

$$\begin{aligned}
Q(\mathcal{M}, p \wedge q) &= Q(\mathcal{M}, p) \cdot Q(\mathcal{M}, q) \\
Q(\mathcal{M}, p \vee q) &= \max(Q(\mathcal{M}, p), Q(\mathcal{M}, q)) \\
Q(\mathcal{M}, \neg p) &= 1 - Q(\mathcal{M}, p) \\
Q(\mathcal{M}, \Box_a \phi) &= \frac{|\{w \in \mathrm{app}_a(\mathcal{M}_0),\ w \models \phi\}|}{|\,\mathrm{app}_a(\mathcal{M}_0)|}
\end{aligned}$$

Now consider what effect agent b announcing that they are a Villager should have on agent a's beliefs. In some worlds that agent a considers possible, agent b might actually be a Werewolf, and agent b surely would like agent a to no longer consider these worlds possible, but since players may lie, agent a can not actually rule out the possibility that agent b is a Werewolf. Indeed, it might seem like the effect of such a statement by agent b is none, since the statement may be the truth, in which case all worlds in which agent b is a Werewolf would have to be discarded, or it is a lie, in which case all other worlds would have to be discarded, but there is no way to know which of the two cases to apply. Instead of eliminating worlds, though, agent a can simply mark them. In this example, in all worlds in which agent b is a Werewolf, despite their statement of the opposite, agent a would mark that if this world was the actual world, agent b would have lied once. As more statements by more agents are made, each world will have a different set and number of agents that must have lied if that world was the actual world.

At this point, we need to consider how an agent can interpret these marks later on. In a way, they tell the agent how many and which agents had to corroborate a particular story for a world to be true. If the majority of other agents were to collaborate to mislead the agent, there would be no way for the agent to determine the truth anymore, so the agent might as well assume that that is not the case. Indeed, in an allusion to Occam's razor, an agent might just assume the world in which the fewest lies were told to be the most likely one to be the actual world. However, since other worlds, even ones in which everyone lied, can never be ruled out definitively, we will use the number of told lies merely as a weight for each world. Rather than determining the quality of a belief as defined above, we therefore use the ratio between the sum weights of the worlds in which a statement holds and the sum of weights of all possible worlds. These weights are defined as $\frac{1}{1+f}$ where f is the number of lies told in that world. We call this modified measure the *weighted quality* $W(\mathcal{M}, \phi)$ of a belief.

Planning Communicative Actions

One of the key challenges of our domain is the fact that players have a large number of possible actions they can use in any order, and that they can repeat what they said, resulting in a large search space to consider. However, ideally we want our agent to pursue a goal like "Convince the other players that you are a Villager", which is technically impossible to achieve, because the other players may not believe anything the agent says. Rather than defining a binary notion of reaching a goal or not, we therefore use the weighted quality defined above to define how closely a goal is reached, and we provide the agent with several different goals that they can all try to reach simultaneously.

The planning problem our agent is solving therefore takes the current (epistemic) state of the world, a set of (communicative) actions the agent can perform and a set of goals the agent can consider as inputs and produces the plan that most closely reaches any of these goals, and the corresponding goal. We perform this planning operation by assigning each state a value equal to the best weighted quality among all available goals divided by the logarithm of the distance from the start state, to force the agent to perform some exploration of the search-space rather than pursuing a greedy approach that would degenerate into depth-first search in most cases. To be able to use this agent in an interactive context in the future, we want to be able to limit how long it deliberates. Since the plans generated by this process only represent the agent approaching a goal without actually reaching it, we can terminate the planning process after it has explored any number of states and use the best plan found at that point, at the potential expense of optimality. In the general case, the agent would need to exhaust the entire search space to be guaranteed to find the optimal solution, but we will show below that, for the One Night Ultimate Werewolf domain, cutting off the search process after it explores a relatively low number of states does not impact performance in the game.

Our Werewolf agents will start most games with a goal of convincing the other players that they are a Villager, because that has the highest probability of succeeding since there are two Villager cards in the game. To achieve that goal, a typical plan our agent comes up with is: `claimRole(Villager)`, `claimRole(Villager)`, `claimRole(Villager)`, i.e. they will claim to be a Villager on every one of their turns. Repeating the same statement has the effect of reinforcing it, which in real games is often heard with some embellishment like "What can I say? I *really* am a Villager". However, if another player claims to be the Seer and that they have seen the Werewolf player's card, the Werewolf might change plans and perform an action like `claimRole(Seer)`, and thereby claiming to be the Seer themselves, to make the other player's story less believable.

Levels of Commitment

Once our agent has found the best plan for the current situation, i.e. the plan that brings it as close as possible to some goal, it has to decide whether to continue pursuing whichever plan it had before or change to that new plan. The two extremes for how to handle this decision are agents that never drop a plan they started, called *fanatical* by Cohen and Levesque, and agents that make no commitment at all to any

[4] The quality measure is actually defined on arbitrary epistemic logic formulas, but any formulas that don't contain any box operator have quality 1 if they are true and 0 otherwise

plan and always change to the currently best plan, which we call *capricious* agents. Note that the new best plan may be the same as the already existing one, for example when no new or relevant information was obtained since the last time the agent computed a plan. What we are interested in is the range of behaviors between these two extremes. We do this by performing a comparison of the score of the current plan and the new plan, where the score $S(\mathcal{M}, p)$ of the plan p starting at the epistemic state $\mathcal{M}$ is the weighted quality of the goal g that plan is supposed to work towards, $W(\mathcal{M}, g)$. An agent that has a current plan p that was formed in some previous state $\mathcal{M}$ will switch to a new plan q computed in the current state $\mathcal{N}$ if and only if

$$S(\mathcal{N}, q) > \alpha \cdot S(\mathcal{M}, p)$$

where we call α the *level of commitment*. A value of 0 for α will cause the inequality to hold for any non-zero score of the new plan, thus prompting the agent to always switch plans, corresponding to a capricious agent, while $\alpha = \infty$ would mean that the agent will never change plans, corresponding to a fanatical agent. Values between these two extremes control how often an agent will change plans, with values between 0 and 1 meaning that an agent will change plans if the new plan they computed is not significantly worse than the old plan, e.g. for $\alpha = 0.5$ the agent will change if the new plan is at least half as good in the current state as the current plan was in the previous state. Values of α that are greater than 1 correspond to agents that only change if significantly better plans are found given the new information they have obtained.

Evaluation

To demonstrate that agents perform differently when they have different levels of commitment to their goals, we first defined goals for the agents to pursue over the course of the game. These goals are at a very high level, to make it necessary for the agents to calculate different plans depending on the situation. They are:

- If the agent is reasonably certain that they are still the Werewolf, they will try to convince the other players that they are some role from the Millers Hollow team.
- If the agent is reasonably certain that they know who the Werewolf is and that they are not one themselves, they will try to convince the other players that that player is a Werewolf.
- If the agent is reasonably certain that they are some non-Werewolf role, they will try to convince the other players that they are that role.

As can be seen, the roles correspond to the general idea of the game of the Werewolves trying not to be detected and the citizens of Millers Hollow to find the Werewolves. Also note that the definition of "reasonably certain" is variable. For most of our experiments we used a weighted quality of 70%, but we will describe below how varying this value changes the behavior of the agents.

The actual evaluation was done through several different experiments. For each experiment, we set up the game by assigning different roles to each of the 5 players, having each player be controlled by a certain type of AI agent and running 4000 games with the same settings. In each of our scenarios, player A started as the Werewolf, there was no other Werewolf present in the game, and we assigned AI agent types by faction, i.e. player A, being on the Werewolf team, was controlled by one type of AI, while players B, C, D, and E, who constituted Millers Hollow, were controlled by four agents of another type of AI. After 4000 games, we measured the percentage of games that player A won, and recorded it as the win rate of the AI agent type that controlled player A versus the AI agent type used by the other four players. Note that player A may win when the Millers Hollow faction wins if their card was exchanged with another player's card. With 4000 games the reported win rate is accurate to $\pm 1.5\%$ in the 95% confidence interval.

Finally, also note that all agents choose their actions at random during the night phase, like which card to look at for the Seer, or whether to swap cards for the Rascal, and only plan their actions during the day phase. Since no player has any information about the other players at this point of the game, this does not logically differ from actual game play behavior. At the conclusion of the day phase all agents use the same strategy again and vote for the player they think is most likely to be a Werewolf other than themselves. We do this to be able to compare the result of different agent types on the task of selecting communicative actions.

State Space Coverage

As described above, the planning problems our agents have to solve often don't have exact solutions, but we are able to terminate the search at any point after the agent has visited n different states and simply use the best plan found at that point. To verify that our agent actually covers enough of the state space to perform well, we ran an experiment in which we compared the performance for different values of n to each other. This experiment showed that the agents' behavior was degenerate, leading to 100% win rates for one side, when very few states are expanded, but when the agents could expand $n = 25$ or more states the win rate was in the expected range. Adding more visited states also did not change the win rate in a statistically significant way. To be very conservative, we let agents expand up to 100 states for the remainder of our experiments.

Levels of Commitment

In this section we will describe the main experiments we performed. We started by establishing a baseline where agent commitment is not relevant to the outcome of the game. We then consider a game that can be considered more typical, in which there are some opportunities for changing allegiances, but not an overabundance of information. This is followed by a setup in which Millers Hollow has a strong information advantage over the Werewolf, to see how a Werewolf player can cope in such a lopsided game. Because the Werewolf player struggled in this scenario, as predicted, we also provide a follow-up experiment that changes how suspicious they are to investigate what effect that has on the outcome of the game.

Baseline: As a baseline we considered an agent that chooses their communicative actions completely randomly, as well as the capricious agent that chooses a new plan to pursue every turn, since it does not commit to any plan. As an initial experiment and basic verification, we set up a game in which player A was the Werewolf and would remain so for the remainder of the game because there was no role present that could change that. Because of this static assignment, we expected the level of commitment to a plan to have little to no effect on the win rate of player A, which was supported by the results of our experiment: For any combination of capricious and balanced agents the win rate for the Werewolf was about 50%, the win rate of a balanced Werewolf player against random agents for Miller Hollow was about 90% for all cases, and the win rate of a random Werewolf against Millers Hollow controlled by any balanced agent type was about 15%, with no statistically significant difference found in any of the cases.

Typical game: As described above, in the case where the assignment of the Werewolf player does not change over the course of the game, different levels of commitment played no significant role in the outcome of the game. However, our hypothesis was that the level of commitment impacts performance in games in which information revealed during game play actually requires players to change their strategy. To test this we performed an experiment in which player B was the Rascal and player C an ordinary Villager. In this scenario, the Werewolf may be exchanged with the Villager, and the Werewolf player has to determine when it is beneficial to change strategies. Figure 1 shows how Werewolf agents with several different levels of commitment perform against the random and capricious baselines. As can be seen, the level of commitment of the Werewolf player influences their win rate, with a commitment of $\alpha = 1$ performing significantly better against the random agents, but performing significantly worse against the capricious baseline. While this result supports our hypothesis that the level of commitment has an effect on agent performance, it also showed that the magnitude and, more importantly, direction of that effect depends on the strategy of the opponent. Additionally, we also performed the same experiment with the AI types reversed, but the level of commitment of the Millers Hollow players had no significant effect on the outcome, with the win rate for a capricious Werewolf player always being close to 50%.

Lopsided game: We also set up the game in a way that makes it a lot harder for player A to win, by assigning the Rascal to B, the Insomniac to C, the Robber to D and the Seer to E. This means that all other players have a huge information advantage over player A, it is unlikely that player A will stay the Werewolf, and if they are not, the player that ends up with the Werewolf will know that they have it because they're either the Insomniac or the Robber. Against a capricious Millers Hollow, player A only wins in around 4% of the games, with no significant difference between the different levels of commitment. However, the interesting behavior in this experiment was that the level of commitment of the Millers Hollow players changed the Werewolf win rate, with Millers Hollow players that are very committed to its plans (with $\alpha \geq 2$) actually dropping the Werewolf win rate to 0% against a capricious Werewolf. Because of this abysmal performance of the Werewolf player, we also ran Millers Hollow populated by balanced agents with $\alpha = 2$, as well as Millers Hollow populated by fanatical players (i.e $\alpha = \infty$), against a balanced Werewolf with several different values for α ranging from 0.1 to 2, but the win rate for player A remained at 0% in all cases.

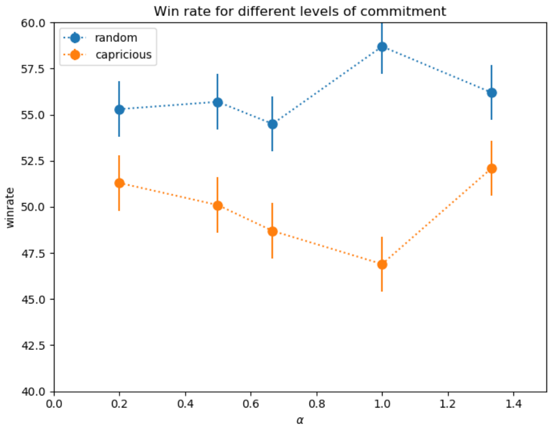

Figure 1: Werewolf win rate for different levels of commitment α against Millers Hollow players controlled by the random and capricious baselines. Note that the graph is scaled to show win rates between 40% and 60% to make the change in win rates visible.

Suspicious Werewolf: The problem in this case is that if the agent is over-committed, it is unlikely to change its plan even when the information revealed by the other players would indicate that the player is no longer on the Werewolf team. However, in addition to commitment to a plan, the certainty agents require of their perception of the world also changes when they change plans. As stated above, by default our agents will assume that they are a Werewolf if that fact has a weighted quality of 0.7 in their mental model of the world. By lowering this required quality, we can make our agents more gullible. For this particular scenario, lowering the required weighted quality to 0.5 resulted in a win rate of 4% for a balanced Werewolf against a fanatical Millers Hollow. Lowering the quality even more had no additional effect on the performance, but resulted in several scenarios in which a Werewolf would volunteer the information that they are a Werewolf because they didn't believe it with sufficient certainty.

Conclusion and Future Work

We presented a novel approach to intentional agents using a quality measure derived from a possible worlds model for agent beliefs. We show how our agents can play a variant of the game One Night Ultimate Werewolf in which the communicative actions are modeled using Dynamic Epistemic Logic. Our agents have an adjustable level of commitment to the goals they pursue, which we hypothesized to have an effect on the outcome of the game. We performed several experiments to verify this claim and presented their results.

While our work investigates how different AI agents perform when playing against other agents, we intend this as the basis for agents that play against human players. There are some key differences that have yet to be addressed, though. The work presented in this paper focused strictly on the logical aspects of the game, but play with human players requires a conversational component. For example, our agents will never ask other agents what their role is, or try to be suspicious of players that withhold information. Another key difference is in what we want to measure. The success in games with humans is better characterized by players' enjoyment of the game rather than raw win rates.

Finally, while our work addresses how committed agents are to their plans, this commitment level has to be defined at the beginning of the game, and is static over the entire duration of the game. However, as our results show, the optimal commitment level depends on the opponent's strategy. It would therefore be advantageous to determine how committed an agent should be to a plan dynamically, and potentially even change it over the course of the game.

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

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.

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.

De Weerd, H.; Verbrugge, R.; and Verheij, B. 2013. How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence* 199:67–92.

Dennett, D. C. 1971. Intentional systems. *The Journal of Philosophy* 68(4):87–106.

Eger, M., and Martens, C. 2017. Practical specification of belief manipulation in games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 30–36.

Eger, M.; Martens, C.; and Cordoba, M. A. 2017. An intentional AI for hanabi. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 68–75.

Gallagher, H. L.; Jack, A. I.; Roepstorff, A.; and Frith, C. D. 2002. Imaging the intentional stance in a competitive game. *Neuroimage* 16(3):814–821.

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.

Mateas, M. 2003. Expressive AI: Games and artificial intelligence. In *DiGRA - Proceedings of the 2003 DiGRA International Conference: Level Up*.

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 (CT'97)*, 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 *IEEE Symposium Series on Computational Intelligence*, 1–8.

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

Shirvani, A.; Ware, S.; and Farrell, R. 2017. A possible worlds model of belief for state-space narrative planning. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 101–107.

Teutenberg, J., and Porteous, J. 2015. Incorporating global and local knowledge in intentional narrative planning. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1539–1546. International Foundation for Autonomous Agents and Multiagent Systems.

Thorne, B., and Young, R. M. 2017. Generating stories that include failed actions by modeling false character beliefs. In *Working Notes of the AIIDE Workshop on Intelligent Narrative Technologies*.

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.