An NLP Interface for Social AI Agents in The Resistance: Avalon

Harrison West
Baylor University
harrison_west1@baylor.edu

Matthew Fendt
Baylor University
One Bear Place 97356
Waco, TX 76798
matthew_fendt@baylor.edu

ABSTRACT

Social deduction games such as Avalon present a unique challenge for AI agents. To discover the hidden roles of others, players must employ indirectness and deception in their communication. DeepRole, an Avalon-playing AI agent created by MIT researchers in 2019, can communicate through in-game actions but is unable to communicate in natural language. We have created Avalocution, a bot that enhances DeepRole with one-way bot-to-human natural language utterances informed by DeepRole's internal knowledge representation. We hypothesized that our natural language interface would produce direct and indirect communication, exhibit human-like behavior, and provide a positive gameplay experience for human players. We collected survey data from research participants who played Avalon against Avalocution agents, and the survey data supports our hypotheses. We conclude that adding Avalocution's simple one-way utterance generation model to DeepRole's existing decision-making framework captures the nuance of communication required in Avalon while providing an excellent gameplay experience.

Keywords

Natural language processing, social deduction games, artificial intelligence

INTRODUCTION

Consider these three statements:

(1) "It's important to maintain a strong team dynamic, and I believe this selection will lead us to victory" (Shi et al. 2023)
(2) "Reviewing past mission outcomes, it's worrisome that Player 4 was involved in a failure. Thus, reassessing our team composition for the next quest may be sensible" (Wang et al. 2023).
(3) "Their combined skills and perspectives can greatly benefit our mission, and I urge all players to consider this team thoughtfully for the betterment of our cause. Let us unite our strengths and work together seamlessly to overcome any challenges that may arise" (Light et al. 2024).

Proceedings of DiGRA 2025

These are direct quotes from bots that play *Avalon*, the social deduction game that is the subject of this research. It is easy to tell that they are LLM-generated: they are clear in meaning, but they lack nuance and are too formal. Few humans would produce statements like these while playing a game.

This is how AI assistants such as ChatGPT generally communicate, and for most tasks, users do not seem to mind. If LLM bots are designed to play cooperative games with humans, perhaps they could be instructed to adopt a more brief and informal style. Of course, their communication would still lack nuance and subtlety— LLMs struggle at poetry and creative writing—but for most games, even for most social games, that should not be a hindrance to gameplay. *Avalon* is different. Avalon requires players to employ indirect communication: to convey what they mean without actually saying it. This challenge arises routinely in the real world but rarely in games, and rarely is artificial intelligence challenged with it.

Scientific interest in games that parallel real-life situations is growing (Bailis et al. 2024; Braverman et al. 2008; Jaderberg et al. 2019; Serrino 2019), and few games parallel real life the way *Avalon* does. It is an eye-opening challenge both for humans and for AI, which has excelled at simpler games for a long time, but which has only recently started to improve at social deduction games such as *Avalon*. To understand the unique challenge that *Avalon* presents to humans and AI agents, we first need to consider other challenges and where they fall short.

BACKGROUND AND RATIONALE

In two-player games of perfect information, AI tends to vastly outperform humans. In 1997, IBM's Deep Blue defeated world chess champion Gary Kasparov in a 6-game match (Campbell et al. 2002), and in 2018, DeepMind's AlphaZero achieved unprecedented performance in Chess, Go, and Shogi (Silver et al. 2018). AI continues to improve at imperfect information games (IIGs) such as poker. In 2018, an AI called Libratus surpassed human professionals in heads-up no-limit Texas hold'em (Brown and Sandholm 2018), and in 2024, researchers created PokerGPT, an LLM-driven AI that efficiently obtained competitive winrates against multiple opponents (Huang et al. 2024). AI can even communicate and negotiate to achieve their competitive goals. The board game *Diplomacy* (Allan B. Calhamer 1959) requires players to communicate in natural language, forming and betraying alliances with other players. In 2022, Meta's AI agent CICERO achieved human-level performance at this especially difficult task (Bakhtin et al. 2022). AI's success at collaborating with humans is even more impressive, as ChatGPT and other GenAI chatbots have transformed the way humans work. AI has also learned to cooperate and compete at the same time. In 2024, researchers introduced an agent (Sidji et al. 2024) that collaborated with humans in the two-team board game *Codenames* (Czech Games Edition 2015). Cooperative and communicative AI face the additional challenge of communicating in a human-like way.

AI continues to get better at IIGs, 3+ player games, and games that are neither strictly cooperative nor strictly competitive. However, even in team games like *Codenames*, AI agents know who their teammates are, and can employ simple communication strategies without losing the support of their teammates. Games like *Diplomacy* require shrewd and precise communication, but only one player can win, and rational players only help others to help themselves. What about games where players need to figure out who is on their team? To persuade someone that it is in their own best interest to temporarily ally with you is difficult, but to persuade someone that you are actually on their side is a different task entirely. To challenge AI with this task, we turn to social deduction games.

Social Deduction

There is no consensus on what constitutes a social deduction game (SDG), and few have attempted to rigorously define the genre. SDGs are a subcategory of hidden role games, where players try to discover the hidden roles or allegiances of other players. Due to their emphasis on bluffing, certain hidden role games such as *Coup* (Indie Boards & Cards 2012) are often considered SDGs. They are strictly adversarial, however, and bear stronger resemblance to other adversarial bluffing games such as *Diplomacy* or Poker than they do to true SDGs such as *Avalon*. We propose that SDGs must meet the following criteria:

(1) There are two or more teams consisting of any number of players.
(2) The composition of the teams is not common knowledge.
(3) One or more players is incentivized to discover the composition of the teams.
(4) Players' in-game actions signal their team affiliation.

A feature of most SDGs is free public communication: players may say anything they want as long they say it publicly. This is true of *Werewolf* and *Avalon*, the two games we will consider in detail.

Werewolf / Mafia

Werewolf, also known as *Mafia* (Dimma Davidoff 1986), is often credited as the first social deduction game. A few players are selected as werewolves and the remaining players are villagers. Werewolves know who each other are, while villagers do not. The game alternates between the night, when the werewolves secretly choose to eliminate one player, and the day, when everyone discusses and votes publicly to eliminate a player they suspect is a werewolf. The paradigm of an evil, informed minority against a good, uninformed majority is replicated in numerous SDGs.

Werewolf has been the subject of much academic research. Braverman et al. (2008) determined the optimal strategies for villagers and werewolves depending on the number of detectives, players on the village team who can discover the allegiance of another player once per night. They were interested in the correlation between information and power, as well as the parallels between *Werewolf* and real life. Today, it is the most extensively studied SDG: in 2023, an annual tournament to create the best *Werewolf*-playing AI reached its fifth year (Kano et al. 2023).

Werewolf is an excellent game for studying multiplayer environments where there is ambiguity about whom to cooperate with. It poses a unique challenge to AI agents because it requires strategic play like chess, shrewd communication like Diplomacy, and a unique ability to distinguish friends from enemies. However, there is another level of complexity that *Werewolf*, in its basic form, fails to capture: the complexity of indirect communication. Here, we turn to *Avalon*.

Avalon

Avalon is a 5-10 player social deduction game created by Don Eskridge and released in 2012. Like *Werewolf*, it is played between an uninformed majority, the resistance, and an informed minority, the spies.[1] Gameplay consists of five missions, and each mission consists of a series of public proposals and votes to determine the players who will go on the mission. Any spy who goes on a mission has the option to privately "fail" it, and if three of the five missions fail, the spies win. The resistance, then, must identify the spies and use their numbers advantage to vote against mission proposals that contain spies. If three of the five missions succeed, the resistance wins.

In addition to spies and resistance, *Avalon* requires two special roles: Merlin and the Assassin. Merlin is a resistance who knows who the spies are but must remain hidden. If the resistance manage to succeed three missions, the spies still win if the Assassin correctly guesses Merlin. Merlin, therefore, cannot simply announce who the spies are, because they would discover his identity. Instead, he must use indirect communication. We define direct and indirect communication as follows:

- Definition 1: Direct communication. In direct communication, meaning is clearly stated in the utterance. For example: “I hate this pick.”
- Definition 2: Indirect communication. In indirect communication, meaning is suggested in the utterance but not directly stated. For example: “I’m not sure what I think about this pick.”

Direct communication is necessary in all SDGs, including *Avalon*. Sometimes, the bad guys are obvious, and need to be called out directly. But *Avalon*’s emphasis on indirect communication distinguishes it from many SDGs. To illustrate the difference between *Avalon* and other SDGs, consider how communication among villagers works in basic *Werewolf*. Its primary goal is to distinguish the villagers from the werewolves. Under the social pressure of prolonged lying, werewolves may unintentionally expose themselves with suspicious chat and votes. Another goal is to share opinions about who is a werewolf and who is a villager. Players usually support their opinions with facts and logical reasoning and communicate them directly. There are reasons why a villager might communicate indirectly: out of genuine uncertainty, for example, or to avoid giving werewolves a definite position.[2] But in general, villagers have little incentive to disguise their beliefs or to decipher the beliefs of others—since villagers have no private information, no one’s opinion is inherently credible.

Without special roles, communication in *Avalon* would be similar to communication in *Werewolf*. The addition of Merlin and the Assassin propels indirect communication to the forefront. Merlin, of course, must disguise his knowledge to protect his identity, and likewise, basic resistance must share their opinions discreetly, both to imitate Merlin and to avoid hurting their Merlin candidacy by saying something wrong. Reading people’s allegiance is central to *Werewolf*; reading people’s beliefs is central to *Avalon*.[3] Opinions in *Werewolf* must be based on logic and facts; in *Avalon*, completely illogical opinions still deserve credence because Merlin could be the one saying them. Human and AI players alike must learn to communicate indirectly and to decipher the indirect communication of other players.

DeepRole

DeepRole (Serrino 2019) is an *Avalon*-playing AI agent trained through self-play using contrafactual regret minimization. Zinkevich et al. (2007) introduced counterfactual regret minimization (CFR), a technique for computing approximate Nash equilibria in extensive-form games, particularly IIGs. The approach decomposes overall regret into counterfactual regret, which is defined for each player at each information set I as the difference in expected utility between the action taken and the best alternative action, conditioned on I being reached and assuming the player had acted to reach it. Contrafactual regret is also weighted by the probability that I would be reached under the current strategy profile, excluding the player's own contribution. Minimizing counterfactual regret independently at each information set leads to a minimization of overall regret, enabling convergence to a Nash equilibrium in self-play.

CFR has been applied extensively to poker. Zinkevich et al. (2007) created a CFR bot that outperformed all its competitors in the 2006 AAAI Computer Poker Competition. Lanctot et al. (2009) introduced chance-sampled CFR, a variant for IIGs that samples a single chance outcome per iteration and updates only the relevant part of the game tree, resulting in less precise but faster strategy updates, a worthwhile trade-off for large games. DeepRole employs chance sampled CFR and CFR+ regret matching (Tammelin 2014), and unlike CFR algorithms for Poker, DeepRole extends the public game tree to be a history of third-person observations rather than player actions alone. This is necessary because unlike Poker, there are non-public player actions in Avalon (succeeding and failing missions).

In 2019, DeepRole was released on the website proavalon.com, where humans had the opportunity to play with it. Despite being trained only through self-play, DeepRole performs well in games with four humans and one DeepRole agent as well as in games with four DeepRole bots and one human. Its biggest limitation, however, is that it does not talk, and can only communicate through picks and votes. Jack Serrino, the creator of DeepRole, says that language is "definitely the next frontier" for bots that play *Avalon* and similar games. In this research, we introduce Avalocution, in which we extend the functionality of DeepRole with an NLP interface, allowing it to produce natural language utterances.

A key aspect of DeepRole's architecture is its public perspective, derived from the public game tree. The public perspective is updated after every move and contains the probability of each possible arrangement of roles from a third-person perspective. It forms the basis for our utterance generation model.

DEVELOPMENT

Avalocution is an NLP interface built on top of DeepRole. It is programmed in Java, and is compatible with DeepRole after modifications to DeepRole's Python code. Given the current state of an *Avalon* game, the public perspective from DeepRole, and the previously conveyed perspective of each bot (see Ranking IGUs), it produces one natural language utterance for each bot. Our goal with Avalocution is to create a pipeline from DeepRole's public perspective to natural language utterances, followed by understanding in the minds of human players.

Class Structure

Our utterance generation model starts with about 100 *base utterance* templates and fills them in with inputs from the current game (such as a speaker and an addressee) to form over 1000 possible *in-game utterances* (IGUs), which can be filtered and evaluated to find the best utterances.

Hidden states and perspectives

Our model relies on *hidden states* and *perspectives*, concepts introduced by DeepRole. A hidden state simply maps each player to a role: Merlin, Assassin, Servant or Minion. In a 5-player game with Merlin and Assassin, there are 60 possible hidden states, but only one of them is the true hidden state. A perspective maps each possible hidden state to its probability of being the true hidden state such that the probabilities sum to 1. DeepRole's public perspective is an example of a perspective. Our model calculates each bot's *private perspective* by taking the public perspective, zeroing out the probabilities of hidden states where the bot is a spy, and scaling the other probabilities back up such that they sum to 1. A bot's utterances are based on its private perspective. A bot's private perspective is generated without regard to the bot's role: if a bot is Merlin or a spy, it knows who the spies are, but it still zeroes out only those hidden states where it is a spy itself. In other words, all bots talk as if they are basic resistance, with no special information.

Base utterances

A base utterance contains immutable, hardcoded information about a statement in *Avalon*. Examples of base utterances and their data are shown in Table 1.

Text	Game State	Intensifiable	Addressee	Conveyed Belief	Conveyed Confidence
“this pick has no spies”	Voting	No	None	Pick has no spies	90%
“X is resistance”	Picking	Yes	None	Target is resistance	85%
“I know you're a spy"	Picking	No	Explicit	Addressee is a spy	90%

Table 1: Three example base utterances, which are used to form in-game utterances (IGUs).

A base utterance has natural language *text*, which may include placeholders for one or more *targets*, players referenced in the utterance. The utterance "X is resistance," for example, requires one target. An utterance's text may include multiple variations with negligible differences in meaning, such as "reject" and "reject this pick." A base utterance has a *game state* during which it is appropriate: either picking or voting, as the bots do not communicate during the mission or assassination phases. A base utterance is *intensifiable* if, when the confidence expressed by its natural language is increased (with an adverbial phrase such as "for sure" or "definitely"), its conveyed confidence in *Avalon* is correspondingly increased. It makes sense to say “X is resistance for sure,” but it sounds a bit awkward to say “this pick has no spies for sure.”

Some utterances, such as "this pick has no spies,” cannot have an *addressee*. Some utterances are required to have an explicit addressee, such as "I know you’re a spy.” Some utterances can only be addressed to certain people. For example, "this is your only team, so you should approve" can only be addressed to someone picked on the current team. Some utterances are implicitly addressed to the mission leader, such as "why would you pick this?" To add variety, these utterances have a chance of including an explicit addressee, as in: "Alice, why would you pick this?"

Base utterances must have at least one *conveyed belief*, and the speaker must be sufficiently confident in the conveyed belief. For example, the utterance "I know you’re a spy” conveys that the addressee is a spy with 90% confidence, and requires the speaker to be 90% confident that the addressee is a spy. A base utterance may also have *preconditions*, which determine if the utterance makes sense in context. For example, the utterance "why repick?" only makes sense if the previous and current picks are identical.

In-game utterances

An in-game utterance (IGU) fills in the attributes of a base utterance with inputs from the current game. A base utterance with the text "X is resistance," could become the following IGU: "Carlos is resistance for sure.”

Algorithm

The Avalocution algorithm runs at the beginning of each picking and voting phase. As mentioned above, it takes three inputs: the game itself, the public belief state produced by DeepRole, and the previously conveyed perspective of each bot. Each bot has an equal probability *p* of producing an utterance, and *p* increases as the number of failed missions increases. Discussion becomes more urgent when the spies are close to winning. For each bot, we convert DeepRole’s public perspective into that bot’s private perspective (see *Hidden States and Perspectives*). Next, given the list of base utterances, we generate an IGU for each legal combination of addressee, targets, and intensification.

From the perspective of one bot, Carlos, in a 5-player game, thirteen IGUs can be generated from just the three utterances in Table 1. “This pick has no spies” is not intensifiable and has no targets or addressee, so it only generates one IGU. “X is resistance” has one target, and any of the other four players can be the target. In addition, the utterance is intensifiable, so Carlos can say “David is resistance” or “David is resistance for sure.” Eight possible IGUs are generated from this base utterance. “I know you’re a spy” has no targets, but any player can fill in as the addressee, producing four IGUs.

Filtering IGUs

After all possible IGUs have been generated, they go through three levels of filtering:

(1) *Are preconditions satisfied?* An utterance such as "why repick?" will be filtered out if the current and previous picks are not identical. Preconditions are evaluated without regard to the speaker's private perspective.

(2) *Is the speaker confident enough?* The speaker must "believe" what the utterance is saying. Suppose that Carlos is 91% confident that David is resistance. As shown in table 1, the utterance "X is resistance," conveys 90% confidence that the target is resistance, which is less than 91%, so "David is resistance" works! On the other hand, the IGU "David is resistance for sure" requires 93% confidence in the same belief. This utterance would be filtered out because Carlos is not sufficiently confident in its conveyed belief.

(3) *Is the speaker too confident?* If the speaker is 100% confident in all the conveyed beliefs of an utterance, that utterance is filtered out. Usually, this step only filters out utterances that are guaranteed to be true based on mission failures. If Carlos and Erin go on a 2-player mission that fails, Carlos will not say "Erin, I know you're a spy" because that is implicitly what he believes.

Ranking IGUs

Utterances that pass the filters are assigned a score based on their *confidence surplus* and *conveyed difference*.

(1) An utterance's *confidence surplus* is the difference between its conveyed confidence and the speaker's actual confidence. In our earlier example, the IGU "David is resistance" has a 1% confidence surplus: the speaker is 91% confident that David is resistance but only conveys 90% confidence. A low confidence surplus is preferred: that way, an utterance more closely approximates the speaker's actual beliefs.

(2) An utterance's *conveyed difference* is the difference between its conveyed confidence and the confidence of the speaker's previously conveyed perspective—the average of the conveyed beliefs of all the speaker's previous utterances, combined with the speaker's implicit belief that they are resistance. A high conveyed difference leads to a better score. That way, bots avoid repeating beliefs they have already conveyed and are quick to disavow beliefs they no longer hold.

Each valid utterance is assigned a score based on these two factors. The utterances are ranked by their score, and an utterance is chosen randomly, with higher-scoring utterances having a greater chance for selection.

Indirect communication

As mentioned earlier, a bot's private perspective is generated without regard to its role. Suppose that Carlos is a spy or Merlin. The examples in this chapter still work. He knows who the spies are, but he still goes through the process of generating IGUs and filtering them out based on his private perspective, which is generated as if he is a basic resistance with no additional information.

In the corpus of base utterances, indirect utterances require less confidence than direct utterances. The direct utterance "this pick has no spies," requires 90% confidence that the pick has no spies, whereas the indirect utterance "wow, interesting pick," requires 30% confidence in the same belief. Regardless of their role, bots naturally produce these indirect utterances: basic resistance due to a genuine lack of confidence, and Merlin and the spies due to a feigned lack of confidence.

EXPERIMENTAL DESIGN

The aim of our experimental design was to evaluate these four hypotheses:

H1: The bots communicate both directly and indirectly. Given *Avalon*'s unique emphasis on indirect communication, we wanted to demonstrate that Avalocution is able to produce both direct and indirect utterances. Direct communication is important in all SDGs. For most games, even many SDGs, indirect communication is unnecessary, but it is a requirement for successful *Avalon* communication.

H2: The bots emulate human behavior with their utterances, picks, and votes. Social interactions with humans are a significant draw of SDGs: humans want to feel like they are collaborating with other humans. Since humans are the gold standard for social deduction teammates and opponents, we wanted to demonstrate that Avalocution produces human-like behavior.

H3: The utterances contribute to a fun and meaningful player experience. Given that games are supposed to be fun and engaging, we wanted to demonstrate that Avalocution utterances lead to a fun player experience. However, we do not want the utterances to just be for fun, we also want them to have a meaningful effect on the gameplay. Sidji et al. (2024) describe how their Codenames-playing AI provided meaningful suggestions that influenced the decisions of human players and also added an extra layer of enjoyment and hilarity to the game. We want Avalocution to do both.

H4: Human players who have played *Avalon* will be more critical of the bots' ability to emulate human behavior. We hypothesized that players with prior *Avalon* experience would scrutinize the bot more heavily, especially in its emulation of human behavior, since new players have no experience of human behavior to which they could compare our bots' behavior. If this hypothesis holds, it limits the effectiveness of Avalocution to people who have little to no experience with *Avalon*.

Experiment

To evaluate Avalocution, we invited human players to play with our bot and answer survey questions about the bot's performance. We recruited 30 undergraduate students to participate in the study. For students in a certain class, participation in the study was one of two possible ways to earn extra credit; other students were personal contacts of the principal investigator and were not in the class, so they did not receive extra credit. Participants reported their own experience with selected social deduction games: 24 had played Mafia or Werewolf, 20 had played Secret Hitler, 18 had played Avalon, 11 had played One Night, 4 had played Town of Salem, and none had played Blood on the Clocktower.

Each participant met in person with the principal investigator to play three games of *Avalon*, with different roles each game: basic resistance, Assassin, and Merlin. The average participant took between 15 and 20 minutes in total to play the three games. During gameplay, the principal investigator was present to answer questions and record qualitative observations about the participants. Afterwards, each participant filled out a survey. Some survey questions were specifically designed to evaluate our hypotheses, while other questions collected background information, identified potential areas of improvement for Avalocution, and determined other factors contributing to a player's experience, such as their satisfaction with the user interface. Two questions concerned direct and indirect communication. At the beginning of the survey, all research participants were given the definitions of direct and indirect communication that we presented earlier.

Hypothesis Evaluation

We asked two survey questions to evaluate each of our first three hypotheses. The questions corresponding to H1 are labeled 1a and 1b, and the questions corresponding to H2 and H3 are similarly labeled.

1a) The bots produced utterances that exemplified direct communication.
1b) The bots produced utterances that exemplified indirect communication.
2a) The bots produced utterances that could have reasonably been produced by a human player.
2b) The bots' picks and votes could have reasonably been done by a human player.
3a) The bots' utterances influenced my picks and votes.
3b) The utterances made my experience more fun.

For these questions, there was no clear comparison group we could use to run a two-population test, so we looked for general agreement to confirm our first three hypotheses. For our fourth hypothesis, however, we did have two populations. We asked research participants if they had previously played *Avalon* or its equivalent games, and divided the data into two populations based on their responses. For questions 2a and 2b, Likert scale responses were binarized, with "Agree" and "Strongly Agree" grouped as "yes" and all other responses grouped as "not yes." We constructed a 2x2 contingency table to compare the frequency of "yes" versus "not yes" responses between the two groups, and we ran one-tailed Fisher's exact test to assess whether the non-players answered "yes" more often. A significance level of α = 0.05 was used.

RESULTS

Responses to the six principal survey questions are shown in Figure 1. Overall, survey data supports hypotheses 1-3. Table 2 shows the percentage of participants who responded "agree" or "strongly agree" to each question. The percentages range from 73% to 90%. These results support our hypotheses that the bots communicate both directly and indirectly, the bots emulate human behavior, and the utterances contribute to a fun and meaningful player experience.

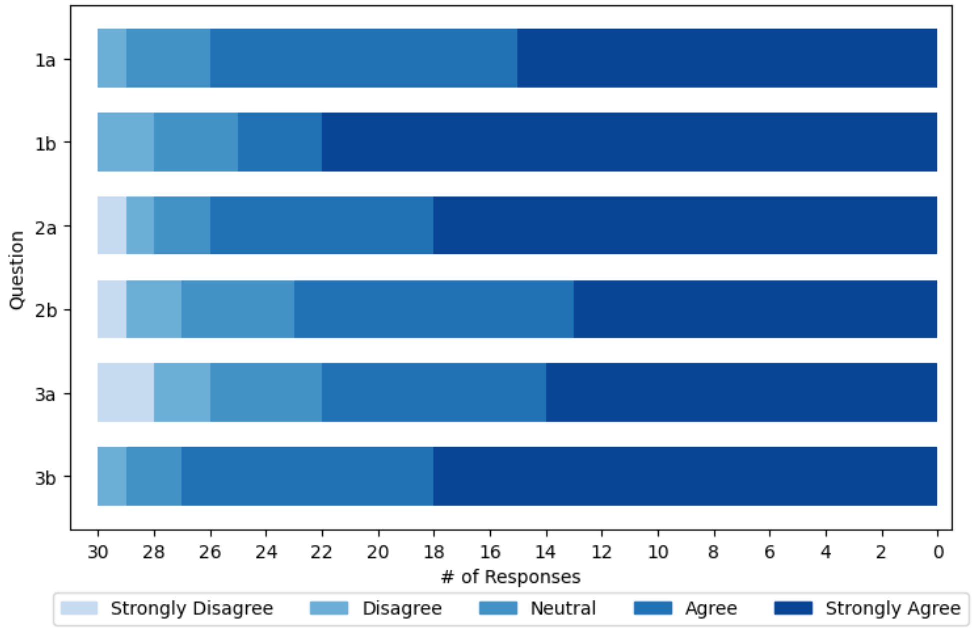

Figure 1: Responses of 30 participants to six principal survey questions after playing three games with our bots. For all questions, the majority of participants responded "agree" or "strongly agree," supporting our first three hypotheses.

Question	Percentage
1a	87%
1b	83%
2a	87%
2b	77%
3a	73%
3b	90%

Table 2: Percentage of research participants responding "agree" or "strongly agree" to six principal survey questions. The results support our first three hypotheses.

Tables 3 and 4 show the contingency table for questions 2a and 2b, split into two populations based on who had previously played *Avalon*. Fisher's exact test on the contingency table for 2a yielded a p-value of 0.47, a statistically insignificant result. Fisher's exact test on the contingency table for 2b yielded a p-value of 0.13, which is closer to but still greater than the significance level of $\alpha = 0.05$. Therefore, we did not find support for our hypothesis that participants who had previously played *Avalon* would be more critical of the bot's ability to emulate human behavior.

		Played *Avalon* before?	
		yes	no
Response	Yes	15	11
	Not yes	3	1

Table 3: Survey responses to 2a, split by who has played *Avalon* before. The results are statistically insignificant.

		Played *Avalon* before?	
		yes	no
Response	Yes	12	11
	Not yes	6	1

Table 4: Survey responses to 2b, split by who has played *Avalon* before. The results are statistically insignificant.

DISCUSSION

The support for our first three hypotheses indicates that we successfully created a pipeline from the belief state of DeepRole bots to utterances, and from utterances to understanding in the minds of our research participants. We were able to capture most desirable metrics for social deduction players: research participants found that the bots were fun, human-like, and influential in the game, and that they exemplified both direct and indirect communication.

Humans are the gold standard for social deduction teammates and opponents, but it can be difficult to organize a group of human players, especially for a 5-10 player game like *Avalon*. Using a simple one-way communication model, our bots simulate human behavior reasonably well, and future work could build on our model to create bots that are even better at fulfilling the role of human players.

We did not find statistically significant support for our fourth hypothesis. Players who have already played *Avalon* have a conception of how humans typically play *Avalon*; therefore, we predicted that those players would scrutinize Avalocution's ability to imitate human behavior. However, both groups reported that the bots produced reasonably human-like behavior. For question 2b, Fisher's exact test had a low but statistically insignificant p-value. Regardless of the p-value, we are less concerned about Avalocution's picks and votes than we are about utterances, since the picks and votes were produced by DeepRole and were not in our control. Overall, these results are encouraging. If our fourth hypothesis had been confirmed, our bot could have been more limited in scope. Because our hypothesis was not supported, we conclude that Avalocution offers a positive player experience, regardless of whether a player has played *Avalon* or not.

Qualitative Data

We collected qualitative data from our survey and from the observations of the principal investigator. We asked participants to indicate what improvements to Avalocution they would most like to see, choosing no more than five from a list of nine options. We also included a free response section where participants could give any additional feedback.

A logic puzzle

Multiple participants characterized their experience with Avalocution as more of a logic game than a social game. One participant expressed disappointment that "when played with bots, the game becomes more deduction and less social." Another wrote that it was "more like a logic game than a social deduction game," which "wasn't a bad thing." Multiple participants noted that it was difficult to play without seeing people's faces. Additionally, the most requested improvement to Avalocution was "games with more than one human and a mix of humans and bots," with 77% of participants selecting that option, and the second-most requested improvement was "ability to understand human utterances and respond to them," selected by 63% of participants. While the participants certainly had a fun experience, it may not have been a *social* experience. Future work could allow our bots to handle two-way communication, at which point we could insert them into games with multiple humans. Reintroducing the social aspect of SDGs is the next step for our bots.

Informal speech

The base utterances were designed to be short, human-like, and informal, and our participants seemed to enjoy the utterances at the extreme of this design. Negative remarks such as "reject this trash" or "are you stupid?" were especially popular with participants. This observation underscores the importance of human-like informalities in social games. Rather than the essay-like dialog produced by LLM-based *Avalon* bots (see *Introduction*), *Avalon*-playing bots should be instructed to adopt a more informal style to mimic the speech of human players.

Limitations

Although we were generally successful and creating a positive experience for players of all experience levels, there are several limitations to our design.

DeepRole

We developed an NLP interface on top of DeepRole, and we did not have control over DeepRole's decision-making. Despite the general agreement that DeepRole has human-like picks and votes, it still exhibits questionable behavior at times. For example, even when the spies are obvious to everyone (for example, when the fourth mission succeeds) the bots that are resistance will often wait until the last possible moment to pick and approve the correct mission. If only one mission has failed, they will even allow another mission to fail. The delay is frustrating for human players. At best, it accomplishes nothing for the resistance—in one instance, DeepRole's creators described this behavior as Merlin "purposefully rejecting missions to seem ignorant," but bots also exhibit this behavior as basic resistance, and a rational Assassin should disregard this behavior since the correct team is already known to everyone. At worst, it can lead to game-losing mistakes, as the probability of a resistance "accidentally" approving a bad pick while delaying the end of the game is non-negligible. As resistance, DeepRole bots occasionally reject the hammer,[4] and as spies, DeepRole bots are far too quick to reveal themselves by rejecting the hammer or by approving teams that they would never approve as resistance. Overall, the deduction abilities of DeepRole bots are very impressive, but these simple errors make them frustrating to play with. DeepRole's habits of rejecting the hammer and unnecessarily delaying the end of the game could be easily fixed with a set of rules that override DeepRole's choices.

Participants

Recruiting participants with no *Avalon* experience gave us insight into how new players respond to the bots. The participants, however, had little experience in general—only one participant had played more than 25 games of Avalon. Many participants struggled with basic strategy. While playing as resistance, for example, some participants approved teams that were mathematically confirmed to contain a spy. These errors did not directly affect our results, but they call into question the players' understanding of the game and their ability to critically evaluate the bots. We also recognize that our survey may be self-selecting, as people who already liked *Avalon* were more likely to sign up for the survey. Our participant pool was also limited to college students at one university. Future researchers should consider surveying a wider variety of participants, such as players who dislike SDGs or the highly experienced *Avalon* players at proavalon.com, where DeepRole was originally tested. Future researchers could also assess participants' knowledge of the game with a short quiz before beginning the games.

Indirectness

Currently, Avalocution bots do not understand how direct or indirect their speech is. They simple communicate their beliefs directly when they are confident and indirectly when they are less confident. But even when a player is highly confident in their beliefs, they nevertheless may want to communicate indirectly. In the future, the bots could use machine learning to learn *when* to communicate indirectly and directly. *Avalon* is a controlled environment for studying indirect communication. In real life, there is virtually no limit to the possible interpretations of an utterance. By contrast, the meaning of all *Avalon* communication can reasonably be reduced to "these hidden states are more likely, and these hidden states are less likely." If a more robust understanding of indirectness were included in Avalocution's interpretable utterance generation model, it could be used as a platform for studying indirect communication both in *Avalon* and in real life.

CONCLUSION

Due to its emphasis on indirect communication, *Avalon* is an especially challenging game for humans and AI agents and captures the nuance of real-world communication in a way that other games do not. We have introduced Avalocution, an NLP interface that extends DeepRole, an *Avalon*-playing AI, with bot-to-human natural language communication. Avalocution uses DeepRole's interpretable belief state to produce natural language utterances. Based on survey data from humans who played with Avalocution bots, Avalocution was successful in key metrics: it communicates directly and indirectly, exhibits human-like behavior, provides a fun and engaging player experience, and is appealing to players regardless of whether they have previously played *Avalon*. These results are encouraging because they show that a simple one-way communication model can reasonably act in place of a human player.

While our participants had a positive experience, many felt that the social aspect of *Avalon* suffered, and expressed a desire for two-way communication and games with a mix of humans and bots. As LLMs improve, these goals are within reach. We have integrated our explainable utterance generation model with DeepRole's robust decision-making; these could be combined with an LLM to enable human-like two-way communication. Irrevocable rules against strictly suboptimal decisions could also be implemented. An *Avalon* bot that combines all these features would be more than the sum of its parts and would provide an even better gameplay experience.

REFERENCES

Bailis, Suma, Jane Friedhoff, and Feiyang Chen. 2024. "Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction." arXiv. http://arxiv.org/abs/2407.13943.

Braverman, Mark, Omid Etesami, and Elchanan Mossel. 2008. "Mafia: A Theoretical Study of Players and Coalitions in a Partial Information Environment." *The Annals of Applied Probability* 18 (3). https://doi.org/10.1214/07-AAP456.

Brown, Noam, and Tuomas Sandholm. 2018. "Superhuman AI for Heads-up No-Limit Poker: Libratus Beats Top Professionals." *Science* 359 (6374): 418–24. https://doi.org/10.1126/science.aao1733.

Campbell, Murray, A.Joseph Hoane, and Feng-hsiung Hsu. 2002. "Deep Blue." *Artificial Intelligence* 134 (1–2): 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1.

Dostoyevsky, Fyodor. 2012. *Crime and Punishment*. Dover Thrift Editions. Newburyport: Dover Publications.

Huang, Chenghao, Yanbo Cao, Yinlong Wen, Tao Zhou, and Yanru Zhang. 2024. "PokerGPT: An End-to-End Lightweight Solver for Multi-Player Texas Hold'em via Large Language Model." arXiv. https://doi.org/10.48550/arXiv.2401.06781.

Jaderberg, Max, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, et al. 2019. "Human-Level Performance in First-Person Multiplayer Games with Population-Based Deep Reinforcement Learning." *Science* 364 (6443): 859–65. https://doi.org/10.1126/science.aau6249.

Kano, Yoshinobu, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, et al. 2023. "AIWolfDial 2023: Summary of Natural Language Division of 5th International AIWolf Contest."

Lanctot, Marc, Kevin Waugh, Martin Zinkevich, and Michael Bowling. 2009. "Monte Carlo Sampling for Regret Minimization in Extensive Games."

Light, Jonathan, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. 2024. "Strategist: Learning Strategic Skills by LLMs via Bi-Level Tree Search." arXiv. http://arxiv.org/abs/2408.10635.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, et al. 2022. "Human-Level Play in the Game of *Diplomacy* by Combining Language Models with Strategic Reasoning." *Science* 378 (6624): 1067–74. https://doi.org/10.1126/science.ade9097.

Serrino, Jack Samuel. 2019. "Finding Friend and Foe in Avalon with Counterfactual Regret Minimization and Deep Networks." MIT.

Shi, Zijing, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. "Cooperation on the Fly: Exploring Language Agents for Ad Hoc Teamwork in the Avalon Game." arXiv. http://arxiv.org/abs/2312.17515.

Sidji, Matthew, Wally Smith, and Melissa J. Rogerson. 2024. "Human-AI Collaboration in Cooperative Games: A Study of Playing Codenames with an LLM Assistant." *Proceedings of the ACM on Human-Computer Interaction* 8 (CHI PLAY): 1–25. https://doi.org/10.1145/3677081.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (6419): 1140–44. https://doi.org/10.1126/science.aar6404.

Tammelin, Oskari. 2014. "Solving Large Imperfect Information Games Using CFR+." arXiv. https://doi.org/10.48550/arXiv.1407.5042.

Wang, Shenzhi, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. "Avalon's Game of Thoughts: Battle Against Deception through Recursive Contemplation." arXiv. http://arxiv.org/abs/2310.01320.

Zinkevich, Martin, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. "Regret Minimization in Games with Incomplete Information."

ENDNOTES

[1] Avalon is a close variation of The Resistance (2009), with an Arthurian re-theme and special roles such as Merlin and Assassin. The terms "resistance" and "spy" come from The Resistance; the equivalent Avalon terms are "loyal servant of Arthur" and "minion of Mordred."

[2] Werewolves can benefit if they know exactly what the villagers are thinking. *Crime and Punishment* character Porfiry Petrovich explains why: "If I shut him up too soon—even though I might be convinced he was the man, I should very likely be depriving myself of the means of getting further evidence against him. And how? By giving him, so to speak, a definite position, I shall put him out of suspense and set his mind at rest, so that he will retreat into his shell" (Dostoyevsky 2012).

[3] *Hanabi* (Antoine Bauza 2010) is a strictly cooperative game that also captures this paradigm. It been studied extensively by AI researchers (Walton-Rivers et al. 2018). It is not an SDG, since there is only one team, and it does not allow natural language communication, but it is a great environment for studying how players signal their private knowledge.

[4] The fifth pick of a mission. If the majority of players reject, the spies win.