Towards Socially Intelligent Agents: An LLM-MARL Framework for Social Deduction Games

Jaywoong Jeong
KAIST
Daejeon, Republic of Korea
jaywoong.jeong@kaist.ac.kr

Abstract

Social deduction games (SDGs) are natural testbeds for evaluating how AI agents reason about others. However, most existing benchmarks are limited to single-game, scripted scenarios that restrict adaptive learning. While Multi-Agent Reinforcement Learning (MARL) suits such multi-party dynamics, integrating open-ended language and social strategy remains challenging.

We propose a unified MARL framework that formulates diverse, language-based SDGs within a shared POMDP. A dense, Theory-of-Mind-inspired reward models how agents influence each other's beliefs—encouraging persuasion for "honest" roles and deception for "hidden" ones. This early-stage work aims to establish a foundation for training and evaluating transferable social reasoning skills such as bluffing, cooperation, and misdirection across different games.

Introduction

Large Language Models (LLMs) show strong performance on complex reasoning and knowledge-intensive tasks (Qiao et al. 2023). However, such successes do not directly translate to social reasoning—the capacity to understand, influence, and coordinate with other agents under partial information. As a result, games have become a practical testbed for evaluating agentic capabilities, from strategic planning to game-theoretic behavior. Social Deduction Games (SDGs) have become a standard testbed for these capabilities, as they inherently require complex social reasoning, with numerous benchmarks developed for games like *Werewolf/Mafia* (Bailis, Friedhoff, and Chen 2024; Kim 2024), *Avalon* (Light et al. 2023), and *Diplomacy* (Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. 2022).

Prior work on SDGs largely relies on static evaluations focused on single games or simplified scenarios (Wang et al. 2025; Kopparapu et al. 2022). These approaches often fail to generalize across different multi-agent contexts and lack mechanisms for active training. While Multi-Agent Reinforcement Learning (MARL) offers a robust framework for training in such settings (Leibo et al. 2021; Kopparapu et al. 2022), effectively integrating natural language remains a significant challenge, with only nascent progress in specific social deduction domains such as *Among Us* (Sarkar et al. 2025). Building on emerging trends in specialized training environments like *NegotiationGym* (Mangla et al. 2025), we propose a common platform that integrates diverse SDGs into a MARL framework, enabling systematic measurement and training of generalized social capabilities.

We address this gap by proposing a unified platform that integrates multiple SDGs within a common MARL formulation. This work presents an early-stage research proposal toward building such a platform; we focus on articulating the formulation, reward design, and evaluation protocol rather than delivering a full implementation. Our contributions are threefold:

1. A unified environment for language-mediated SDGs with Gym-compatible interfaces;
2. A Theory-of-Mind–motivated social reward for persuasion and deception; and
3. An evaluation protocol targeting cross-game generalization beyond win-rate metrics.

Methodology

To systematically train and evaluate generalized social skills, we first propose a common formulation that captures the core components of diverse SDGs.

Problem Formulation

We formulate the social deduction environment as a Partially Observable Stochastic Game (POSG) for N agents, defined by the tuple $\langle \mathcal{S}, \{\mathcal{O}^i\}, \{\mathcal{A}^i\}, \mathcal{T}, \mathcal{R}, \gamma \rangle$:

- $\mathcal{S}$ denotes the **global state space**, encompassing all public information (e.g., player status, public votes) and private information (e.g., player roles, hidden cards).
- $\mathcal{O}^i$ is the **observation space** for agent $i \in \{1, \ldots, N\}$. At each step t, agent i receives a private observation $o_t^i \in \mathcal{O}^i$, which typically includes public game events and their own role/cards. Agents must maintain a belief b_t^i over the true state s_t based on their observation history.
- $\mathcal{A}^i$ represents the **action space** for agent i. We define a composite action space $\mathcal{A}^i = \mathcal{A}^i_{game} \cup \mathcal{A}^i_{comm}$, where $\mathcal{A}^i_{game}$ handles discrete game mechanics (e.g., voting, using an ability) and $\mathcal{A}^i_{comm}$ serves as an interface for natural language utterances.

- $\mathcal{T} : \mathcal{S} \times \mathcal{A}^1 \times \ldots \times \mathcal{A}^N \to \Delta(\mathcal{S})$ is the **state transition function**, describing the probability of moving to the next state given the current state and joint actions.
- $\mathcal{R}$ is the **reward function**, where $r_t^i(s_t, \mathbf{a}_t)$ denotes the reward for agent i. This is often sparse (e.g., reflecting only team win/loss) or shaped based on intermediate objectives to facilitate learning.

Categorization of Game Mechanics

We categorize SDGs using two practical lenses: game-theoretic interaction structure and RL modelability via canonical state/action encodings. The first lens characterizes interaction incentives and information flow (e.g., zero-sum vs. mixed-motive, hidden roles, public voting, asymmetric abilities), as shown in Table 1. The second lens parameterizes implementability by estimating state and action dimensions as functions of players, roles, rounds, and phase primitives. These lenses inform our abstractions (role visibility, voting, leader rotation, challenge mechanics) and guide which skills we test. We present the canonical formulas for state/action sizes in Table 2. Balancing implementability and diversity coverage, we select four representatives—Coup, Mafia, Avalon, and Bang!—as our core evaluation set.

Implementation

We plan to implement a Gym-compatible text-based environment that generalizes to multiple SDGs via thin adapters (Sarkar et al. 2025). Observations concatenate the public state and dialogue, and actions use a composite $a_t = (a_{game}, a_{comm})$ that mixes discrete moves with natural language utterances. The scheduler cycles through gameplay, discussion, and voting phases, returning to gameplay if the episode continues. Each target game (Coup, Mafia, Avalon, and Bang!) maps its native steps and events to this loop and to the unified observation/action, with canonical state/action sizes in Table 2.

Reward Formulation for Social Reasoning

The sparse default reward r_g (e.g., ± 1 at game end) is insufficient for training complex linguistic strategies. To address this, we introduce a dense reward r_s based on modeling the beliefs of other agents, inspired by the "Speaking" and "Listening" mechanisms in Sarkar et al. (2025).

Following computational rationality (Oulasvirta, Jokinen, and Howes 2022), we define an agent's objective as maximizing its influence on the beliefs of other players as a practical operationalization of Theory of Mind. Let $b_j^t(q)$ be the estimate of agent j's belief at time t that a proposition q (e.g., "Player X is the Mafia") is true. To compute this reward effectively, we distinguish between training and execution: **during centralized training**, $b_j^t(q)$ is approximated using the critic's privileged access to agent j's internal state or the ground truth distribution; **during execution**, agents rely on a learned belief model of others (inferred from observed action and dialogue histories.

When an "honest" agent i (e.g., Crewmate) makes an utterance a_{comm}^i, its social reward r_s is based on its ability to move other honest agents $j \in Crew$ closer to the truth:

$$r_s^i = \sum_{j \neq i, j \in Crew} \left(b_j^{t+1}(q) - b_j^t(q) \right)$$

This rewards utterances that successfully persuade others towards the correct belief.

Conversely, for a "hidden" agent k (e.g., Mafia), the objective is to maximize confusion. The reward signal is inverted, rewarding the agent for pushing others away from the truth:

$$r_s^k = - \sum_{j \in Crew} \left(b_j^{t+1}(q) - b_j^t(q) \right)$$

This explicit ToM-based reward allows agents to learn sophisticated strategies of persuasion and deception, rather than just game mechanics.

Proposed Evaluation and Future Directions

Our framework enables a systematic evaluation of generalized social skills, moving beyond simple win-rates. We propose a protocol focused on specific social metrics and cross-game generalization.

Metrics for Social Capabilities

We will evaluate agents not only on task completion (team wins) but also on specific social metrics derived from our ToM-based reward formulation:

- **Role Inference Accuracy:** We will measure an agent's belief $b_i^t(q)$ in the true hidden state q (e.g., Mafia's identity) over time. A successful agent should show increasing accuracy as more evidence (dialogue, actions) is observed.
- **Persuasion & Deception Efficacy:** Using the belief changes of other agents as a proxy, we can quantify an honest agent's "persuasion score" (its ability to align others' beliefs with the truth) and a deceptive agent's "deception score" (its ability to misalign them).

Generalization Case Study

The primary goal of our common framework is to test skill transfer. We will conduct a key case study:

1. Train an agent A_{bluff} exclusively on **Coup**, where it learns individual bluffing (a zero-sum, deceptive skill).
2. Train an agent A_{coop} exclusively on **Avalon**, where it learns team-based logical deduction.
3. Evaluate both A_{bluff} and A_{coop} on a novel, mixed-motive game like **Mafia** (zero-shot or few-shot).

This will allow us to quantify if specific social skills (bluffing vs. cooperation) are generalizable or remain task-specific.

References

Bailis, S.; Friedhoff, J.; and Chen, F. 2024. Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction. arXiv:2407.13943.

Kim, M. 2024. GPTs in Mafia-like Game Simulation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, 1–5. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0331-7.

Kopparapu, K.; Duéñez-Guzmán, E. A.; Matyas, J.; Vezhnevets, A. S.; Agapiou, J. P.; McKee, K. R.; Everett, R.; Marecki, J.; Leibo, J. Z.; and Graepel, T. 2022. Hidden Agenda: a Social Deduction Game with Diverse Learned Equilibria. arXiv:2201.01816.

Leibo, J. Z.; Dueñez-Guzman, E. A.; Vezhnevets, A.; Agapiou, J. P.; Sunehag, P.; Koster, R.; Matyas, J.; Beattie, C.; Mordatch, I.; and Graepel, T. 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. In *Proceedings of the 38th International Conference on Machine Learning*, 6187–6199. PMLR.

Light, J.; Cai, M.; Shen, S.; and Hu, Z. 2023. AvalonBench: Evaluating LLMs Playing the Game of Avalon. https://www.semanticscholar.org/paper/AvalonBench%3A-Evaluating-LLMs-Playing-the-Game-of-Light-Cai/9f2ebd1f09990e98049d4ea6df2729b18f97d8ef. Accessed: 2025-08-13.

Mangla, S.; Hokamp, C.; Boylan, J.; Ghalandari, D. G.; Jauhari, Y.; Cassidy, L.; and Duffy, O. 2025. NegotiationGym: Self-Optimizing Agents in a Multi-Agent Social Simulation Environment. urlhttps://openreview.net/forum?id=c2OO0ktqUV. First Workshop on Social Simulation with LLMs.

Meta Fundamental AI Research Diplomacy Team (FAIR)†; Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer, A.; Lewis, M.; Miller, A. H.; Mitts, S.; Renduchintala, A.; Roller, S.; Rowe, D.; Shi, W.; Spisak, J.; Wei, A.; Wu, D.; Zhang, H.; and Zijlstra, M. 2022. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074.

Oulasvirta, A.; Jokinen, J. P. P.; and Howes, A. 2022. Computational Rationality as a Theory of Interaction. In *CHI Conference on Human Factors in Computing Systems*, 1–14. New Orleans, LA, USA: ACM. ISBN 978-1-4503-9157-3.

Qiao, S.; Ou, Y.; Zhang, N.; Chen, X.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; and Chen, H. 2023. Reasoning with Language Model Prompting: A Survey. arXiv:2212.09597.

Sarkar, B.; Xia, W.; Liu, C. K.; and Sadigh, D. 2025. Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning. arXiv:2502.06060.

Wang, H.; Feng, X.; Li, L.; Guo, Y.; Qin, Z.; Sui, D.; and Kong, L. 2025. TMGBench: A Systematic Game Benchmark for Evaluating Strategic Reasoning Abilities of LLMs. arXiv:2410.10479.

Appendix

Table 1: Game-theoretic structures and core strategic skills.

Game	Zero-sum Type	Strategic Skills
Mafia	Zero-sum	deception detection, trust reasoning
The Resistance	Partial zero-sum	coalition reasoning, mission planning
Avalon	Partial zero-sum	cooperative reasoning, role inference
Secret Hitler	Partial zero-sum	asymmetric coordination, deception management
Coup	Semi zero-sum	bluffing, strategic timing
One Night Werewolf	Zero-sum	fast role inference, temporal reasoning
Bang!	Non-zero-sum	spatial threat modeling, hidden alignment
Shadow Hunters	Non-zero-sum	mixed-motive reasoning, partial information combat
Among Us	Zero-sum	spatial reasoning, social deception
Project Winter	Non-zero-sum	resource coordination, negotiation

Table 2: Parameterised state and action space dimensions under a canonical encoding. Symbols: N players; K role types; M missions; t_r team size in round r; $1\{\cdot\}$ is indicator.

Game	State dimension $S(\cdot)$	Action dimension $A(\cdot)$ (per agent / phase)
Coup	$S(N,K) = 2NK + N + N + K + c_{phase}$ (influence, coins, alive, deck, phase)	Active: $A_{act}(N) = 4 + 3(N-1)$ (income/aid/tax/exchange; target actions); Reaction: $A_{react} = 2 + B$ (challenge, allow; block roles).
Mafia	$S(N,K) = NK + N + c_d + v$ (roles, alive, day/night, vote buffer)	Night: $A_{night}(N) = 1\{mafia\}(N-1) + 1\{doctor\}N + 1\{cop\}(N-1)$; Day: $A_{day}(N) = N$ (vote incl. skip).
The Resistance	$S(N,K,M) = NK + \sum_{r=1}^{M}(1+N) + c_{leader}$ (roles, outcomes+votes, leader)	Leader: $A_{prop}(N,t_r) = Nt_r$; Vote: $A_{vote} = 2$; Mission: $A_{mission} = 2$.
Avalon	$S(N,K,M) = NK + \sum_{r=1}^{M}(1+N) + c_{info}$ (Merlin/Percival info)	Resistance ; Final: $A_{assassin}(N) = N$.
Secret Hitler	$S(N,K) = NK + 11 + N + N + 2$ (roles, policy track, term flags, pres idx, deck comp.)	Nomination: $A_{nom}(N) = N - 1 - ineligible$; Vote: $A_{vote} = 2$; Pres. discard: $A_P = 3$; Chan. discard: $A_C = 2$; Exec.: $A_{exec} \in \{N, N, 1, N\}$.
One Night Werewolf	$S(N,K) = (N+3)K + c_{order}$ (players+center cards, night order)	Night by role: $A_{\max} \in \{N, N2, 3, \ldots\}$; Day: $A_{vote} = N$.
Bang!	$S(N) = N + 4N + N + N + N^2 + C$ (alive, roles, HP, hand, distance, in-play)	Active: $A_{act} \approx (playable) + (N-1) + 1$.
Shadow Hunters	$S(N) = 3N + N + B + C$ (faction, HP bins, board, equip/hand)	Turn: $A_{turn} \approx B + (N-1) + C$.
Among Us†	$S(N,G) = 2N + N + 2N + N\lvert\mathcal{T}\rvert + NG^2$ (roles, alive, meetings, tasks, grid)	Free-play: $\{use, report, kill, vent, sabotage\}$; Meeting: $A_{vote} = N$ (N+1 with skip).