

A Declarative Model for Social Deduction Games with Communication

Lucas Araujo Carvalho¹, Geraldo Xexéo¹

¹Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brazil

{lucasac, xexeo}@cos.ufrj.br

Abstract. Introduction: Social deduction games combine hidden roles, asymmetric information, and strategic communication, making them useful environments for studying deception in games and AI agents. **Objective:** This paper introduces SDGM, an early-stage declarative model in which hidden roles, phase-based actions, and communication channels are treated as first-class constructs. **Steps:** We define SDGM's main abstractions, distinguish material state from communicative trace, and illustrate its concrete syntax through a reduced version of Werewolf. **Expected Results:** SDGM is expected to support structured descriptions, future simulation, expressiveness analysis, and deception tracing in multiagent settings.

Keywords General Game Playing, Deception, Social Deduction Games

1. Introduction

Social deduction games (SDG), such as Mafia and Werewolf, are a class of games centered around hidden roles and deductive reasoning. Their signature mechanic is not merely incomplete information, but strategic communication under asymmetric knowledge. Moreover, special roles add another layer of complexity, as they change what players know and how they interpret the actions of others. This combination of conflicting goals, hidden information, and persuasive interaction makes them suitable environments for studying deception [Starace et al. 2026]. Existing General Game Playing (GGP) research and languages primarily focus on general game mechanics, while communication is typically treated as secondary or external to the game description itself.

To address this gap, this paper introduces the Social Deduction Game Model (SDGM), an early-stage declarative model for representing SDG in which hidden roles, phase-based actions, and communication channels are treated as first-class constructs. The current version focuses on defining the model's main abstractions and an initial concrete syntax, while full operational semantics and engine-level execution remain ongoing work. The model is illustrated through a reduced version of *Werewolf*, showing how teams, roles, channels, actions, and phases can be expressed in a compact specification.

2. Related Works

GGP refers to the development of autonomous systems capable of effectively playing a diverse array of games by interpreting formal descriptions of those games at runtime. These games are typically modeled as state machines, where configurations evolve based on player actions and transition functions. While the original Game Description Language

(GDL) [Love et al. 2006] was limited to deterministic scenarios, later extensions like GDL-II [Thielscher 2010] and GDL-III [Thielscher 2017] introduced mechanisms for partial observability and the modeling of players' epistemic states, respectively. Meanwhile, the Ludii [Piette et al. 2020] system adopts a “ludemic” approach, defining games through conceptual units that enhance description efficiency and readability. Despite these advancements, GGP research remains largely focused on structural games such as board and card games, leaving a gap in representing communication-centered games. SDGM addresses this limitation by treating communication channels and spoken interactions as first-class elements (Table 1).

Strategic communication in these environments often revolves around “cheap talk”, messages that carry no immediate cost to the sender. These messages do not directly alter the game state and have no obligation to be truthful, yet they influence the beliefs and intentions of other players [Farrell e Rabin 1996]. In competitive scenarios with asymmetric information, this can lead to functional deception, where an agent intentionally induces false beliefs in opponents to maximize its own rewards [Ward et al. 2023]. SDG, characterized by hidden roles and conflicting goals, serve as ideal laboratories for investigating these dynamics. Recent AI benchmarks in games like Diplomacy and Avalon have demonstrated emergent deceptive behaviors, highlighting the necessity for new modeling frameworks [Starace et al. 2026].

Table 1. Comparison between SDGM and related game description approaches.

Approach	Hidden info	Epistemic state	Channels	Cheap talk
GDL	No	No	No	No
GDL-II	Yes	Limited	No	No
GDL-III	Yes	Yes	Not central	Not central
Ludii	Partial	No	No	No
SDGM	Yes	Yes	Yes	Yes

3. Proposed Solution

This work proposes the Social Deduction Game Model (SDGM), a formal language designed to represent games in which communication is a central mechanic. Unlike traditional Game Description Languages, which focus primarily on observable state transitions, SDGM explicitly models communication acts as part of the game structure. This enables the representation of environments where persuasion, coordination, and deception directly impact outcomes, even when they do not alter the game state.

3.1. SDGM modeling

The SDGM is designed to be modular and extensible, being able to formalize core SDG and expansions. It incorporates both the structural mechanics of the game and the communication dynamics between players. The goal is to explicitly represent interactions that influence the players' internal beliefs alongside mechanical actions that alter the observable game state.

To achieve this, the SDGM views players as actors performing specific actions within the game. In this framework, roles act as character definitions or “capability bundles” that provide actors with unique initial knowledge and specific action sets. Roles

are assigned to teams, and for each team is assigned a win condition. The goal is for the actors to achieve their team's win condition.

The progression of the game is organized into phases, that dictate which actors are currently active and the possible actions they can perform. In summary, the phase declaration must specify who can perform what at that point of the game. The game cycles through the available phases until a team's win condition is achieved.

Simultaneous actions may be performed in a single phase. To ensure deterministic results, the SDGM engine will implement a resolution pipeline that coordinates the resolution of each action effect, based on the action's priority value. This approach allows the definition of priority-based actions, such as the protection from the medic role in Werewolf.

SDGM distinguishes between the material game state and the communicative trace. Mechanical actions, such as killing or voting, may directly modify material predicates, such as player status. Communication actions, in contrast, do not directly change material predicates. They produce messages that are delivered through channels and recorded in communication logs, which can then be used by agents, simulators, or analysis tools to update beliefs, infer deception, or reconstruct information flow. This distinction allows SDGM to represent cheap talk without treating every utterance as a direct state transition in the underlying game world.

Building on this distinction, SDGM treats communication acts as first-class actions. This enables the modeling of "cheap talk" as explicit moves that do not directly alter the material state of the game, but generate communicative events available to players, agents, and analysis tools. These acts are mediated by communication channels that define the information topology of the game, specifying which actors can speak or listen during a particular phase. By distinguishing between public channels, accessible to all active actors, and private channels, restricted to specific groups or factions, the model represents the information asymmetry essential to the genre and supports the analysis of deceptive communication strategies.

3.2. Language Syntax Prototype

The current SDGM syntax prototype draws strong inspiration from the ludemic approach of Ludii. Similar to Ludii, games are described as compositions of high-level conceptual units, promoting readability and reuse. The language adopts a Lisp-like syntax, where constructs are expressed as nested symbolic expressions. However, SDGM extends this paradigm by introducing first-class constructs for communication. SDGM specifications are organized into the following top-level components:

- **Players:** defines the number of players participating in the game. This is specified as a fixed integer, in the same way that the Ludii descriptions do.
- **Setup:** initializes the game state, including role distribution and global properties. For example, the (dealRoles ...) construct assigns roles to players at the beginning of the game, while (properties ...) define state variables such as "status" (e.g., alive or dead), which are later used in actions and win predicates.
- **Teams:** specifies the different factions in the game and their respective win conditions. These conditions are expressed as logical predicates over the game

state, such as counting the number of alive players of a given role, enabling flexible definitions of victory criteria.

- **Roles:** define the capabilities of each player type. They encapsulate action sets, team membership, and initial knowledge. For example, the werewolf role includes knowledge of the identities of all other werewolves from the beginning of the game.
- **Actions:** describes all possible interactions in the game, including both mechanical and communication actions. Each action includes parameters such as type (e.g., select or communicate), aggregation method (e.g., majority or individual), valid targets, and effects. Effects may modify the game state or reveal information to specific players.
- **Channels:** defines the communication structure of the game by specifying who can exchange messages and under what conditions. Channels can represent public discussions or restricted communication, like a private werewolf channel, and are activated depending on the current phase.
- **Phases:** organizes the temporal progression of the game. Each phase determines which actions are available, which players can act, and which communication channels are enabled. For example, a night phase may allow only werewolves and special roles to act with restricted communication, while a day phase enables global discussion followed by collective decision-making actions such as voting.

Taken together, these components act as modular building blocks for describing SDG in a clear and structured way. By organizing the game into well-defined elements such as roles, actions, and phases, SDGM enables complex mechanics to be expressed in a readable and declarative form. This improves not only the ease of authoring game descriptions but also their interpretability for analysis and reuse.

To illustrate the proposed syntax, Figure 1 presents fragments of a reduced version of *Werewolf* with five players and a single wolf. The example is intentionally partial and serves only to demonstrate how SDGM abstractions are encoded in a compact declarative specification.

<pre>(team "village" (win (= 0 (count (players (where (role "werewolf") (property "status" "alive"))))))))</pre>	<pre>(role "werewolf" (team "werewolves") (abilities ("kill" "lynch" "claim" "accuse" "vote_intent")) (knowledge (reveal role (players (where (role "werewolf")) (to self))))</pre>	<pre>(channel "public_chat" (access (players (where (property "status" "alive"))))</pre>	<pre>(action "kill" (type selection) (aggregation majority (on_tie none)) (options (players (where (not (role "werewolf")) (property "status" "alive")))) (effect (target (set (property "status" "dead")))) (constraints (self (property "status" "alive")) (priority 0))</pre>	<pre>(phase "night_actions" (actors (players (where (property "status" "alive")))) (perform ("kill" "investigate")) (next "day_deliberation"))</pre>
--	--	---	--	---

Figure 1. Illustrative SDGM snippets for a reduced version of *Werewolf* with one wolf. The five columns show partial definitions of a team, a role, a public communication channel, the wolf's night-kill action, and the night phase. The snippets are not intended as a complete game specification, but as a compact example of the proposed concrete syntax.

4. Expected Results

SDGM is intended to provide a specialized layer for SDG rather than a replacement for the general formalisms of the current game description languages. Compared to GDL-III,

which provides a general low-level logic for imperfect information games, SDGM offers a higher-level abstraction that simplifies the modeling of more complex hidden-role games. Similarly, while the Ludii system emphasizes compositional reuse through ludemes, SDGM specifically targets the social deduction genre by making private knowledge and communication acts into first-class elements of the model. This specialization allows for a more readable and declarative description of games where the strategic core comes from social inference.

To ensure the model framework meets the requirements for use in future research, the following evaluation criteria will be verified:

- **Expressiveness:** The model should support concise and scalable representations of a wide range of Mafia-like games, from a basic Werewolf to more complex variants. This includes the ability to represent diverse role interactions, communication structures, and phase systems without loss of clarity.
- **Faithfulness:** The semantics of SDGM must accurately capture the intended mechanics and informational constraints of the modeled games. This will be evaluated by comparing simulation outcomes with known analytical results and expected gameplay dynamics described in the literature.
- **Deception Tracing:** The framework should enable structured logging of communication and action events, allowing researchers to trace how information propagates and how deceptive strategies emerge. This includes generating datasets suitable for downstream analysis in areas such as natural language processing and multiagent learning.

5. Conclusion

This paper introduced the Social Deduction Game Model (SDGM), an early-stage declarative model for representing SDG in which hidden roles, phase-based actions, and communication channels are treated as first-class constructs. By distinguishing material game state from communicative trace, SDGM provides a way to represent cheap talk without reducing it to ordinary state-changing actions. The reduced *Werewolf* example illustrates how teams, roles, channels, actions, and phases can be expressed in a compact syntax. Although no formal evaluation has yet been conducted, these examples provide preliminary evidence of the feasibility of the approach.

Future work will focus on defining the full operational semantics, implementing the SDGM Game Manager, and evaluating the model through more complex games such as *Secret Hitler* and *Avalon*. Ultimately, SDGM aims to support simulation and analysis of deceptive communication and belief dynamics in SDG.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

Disclosure about use of AI

During the preparation of this work the author used ChatGPT in order to improve text flow and clarity. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

- Farrell, J. e Rabin, M. (1996). Cheap talk. *Journal of Economic perspectives*, 10(3):103–118.
- Love, N., Hinrichs, T., Haley, D., Schkufza, E., e Genesereth, M. (2006). General game playing: Game description language specification. Technical Report LG-2006-01, Stanford University, Stanford Logic Group.
- Piette, É., Soemers, D. J. N. J., Stephenson, M., Sironi, C. F., Winands, M. H. M., e Browne, C. (2020). Ludii – the ludemic general game system. In Giacomo, G. D., Catala, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., e Lang, J., editors, *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 411–418. IOS Press.
- Starace, J., Tafoya, J., Singh, A., e Soule, T. (2026). Deceptive algorithms in games: A systematic literature review. *Entertainment Computing*, 56:101078.
- Thielscher, M. (2010). A general game description language for incomplete information games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):994–999.
- Thielscher, M. (2017). Gdl-iii: A description language for epistemic general game playing. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1276–1282.
- Ward, F. R., Belardinelli, F., Toni, F., e Everitt, T. (2023). Honesty is the best policy: defining and mitigating ai deception. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.