Explainable Rational Synthesis in Multi-Agent Systems

Alireza Farhadi[1] Mohammad Izadi[2] Jafar Habibi[2] Tobias Meggendorfer[3]

[1]Kish International Campus, Sharif University of Technology, Iran
[2]Computer Engineering Department, Sharif University of Technology, Iran
[3]Lancaster University Leipzig, Germany

ExCoS 2025 Workshop, Glasgow, Scotland

Outline

Introduction: The Challenge

- Coordinating intelligent agents in high-stakes scenarios like autonomous driving requires protocols that are both efficient and explainable.
- We formalize agent interactions using game theory, where agent objectives are specified in linear temporal logic (LTL).
- **Key Challenges:**
  - **Computational Complexity:** Automatically synthesizing Nash Equilibria (NE) is hindered by double-exponential complexity.
  - **Black-Box Nature:** The resulting strategies are often difficult for humans to comprehend and trust.
- **This Work's Solution:** An integrated approach that combines a performant synthesis algorithm with an explainability framework to make rational synthesis practical and transparent.

The Proposed Algorithm

- This paper introduces a novel and performant synthesis algorithm.
- **Core Algorithm Steps:**
  1. **Suspect Game Construction:** First, it solves a series of parity games via a *suspect game* construction to guarantee punishment for any agent that deviates from its strategy.
  2. **Equilibrium Path Extraction:** Then, it applies SAT-based bounded model checking to extract a minimal equilibrium path from a run graph.

- **Implemented Tool (CGES):** The core synthesis algorithm is implemented in a tool named CGES (Concurrent Game Equilibrium Synthesizer). It demonstrates significant performance gains over state-of-the-art methods.

Framework for Explainability

- To move beyond opaque, "black-box" synthesis, we use two main approaches:
- **1. Port Automata (PA):**
  - We use a PA-based connector framework for transparently modeling game and strategies.
  - PAs enhance explainability through intuitive visualization as state-transition diagrams.
- **2. Contrastive Explanation:**
  - We propose an algorithm for interactive, contrastive "why-not" questioning.
  - This allows users to ask questions like, "Why move X instead of Y?".
  - The algorithm analyzes the equilibrium structure to formally answer such queries. Implementation in CGES is future work.

An Illustrative Example

- Agents A_1 and A_2 interact in a concurrent game.
- At s_0, A_1 chooses a or b, while A_2 selects c.
- A_1's objective is $\mathbf{FG}s_1$.
- A_2's objective is $\mathbf{FG}s_2$.

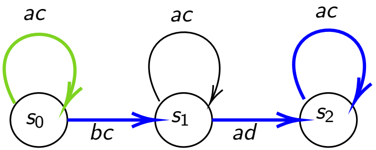
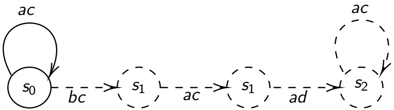


Figure: Concurrent game structure

Example: Nash Equilibria Found

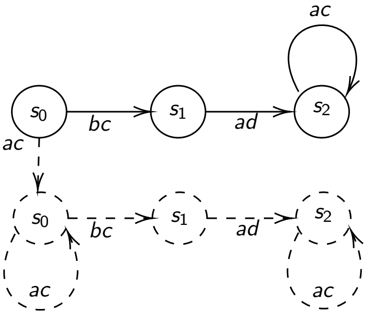
First NE (Green Path)

- In this equilibrium, neither agent achieves its objective.
- If A_1 deviates, A_2 redirects the game to punish A_1.



Second NE (Blue Path)

- In this equilibrium, agent A_1 fails and A_2 succeeds.
- The dashed section neutralizes any futile deviation by A_1.



Evaluation: CGES vs. EVE

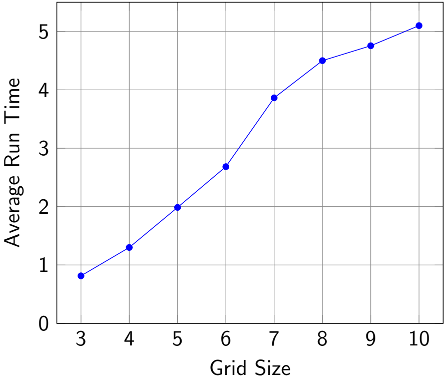
- The CGES tool was evaluated on two case studies and compared with the state-of-the-art EVE tool.
- **Case Study 1: Gossip Protocol**
  - A protocol for information dissemination in large-scale systems.
  - CGES demonstrates significant superiority over EVE in scenarios with larger player counts.

Table: Performance evaluation of EVE and CGES in the gossip protocol

Players	**2**	**3**	**4**	**5**	**6**	**7**	**8**
CGES	0.2s	0.4s	0.6s	1.4s	3.8s	20.1s	167.3s
EVE	0.1s	0.2s	1.1s	13.5s	310.4s	>2 hours	?

Evaluation: Multi-Robot Motion Planning

- **Case Study 2: Multi-Robot Motion Planning (MRMP)**
  - Two robots must navigate an $n \times n$ grid to reach opposite corners without collisions.
  - The average execution time for CGES grows linearly and remains under 5 seconds.
  - In contrast, EVE's performance is reported to be exponential, exceeding 2 hours for a 10×10 grid.



Conclusion and Future Work

Contribution

- This work offers an efficient and explainable rational synthesis method.
- By using suspect games, PA-based visual connectors, and a proposed algorithm for contrastive explanations, we make multi-agent system protocols more transparent and user-amendable.

Future Work

- Implementation of the contrastive explanation algorithm in the CGES tool is planned for future work.
- Future work will also focus on integrating these formal methods with large language model-based agents to guide their strategic decision-making.

Questions?

Thank you for your attention!

Questions and Discussion

CGES Tool: `https://github.com/incaseoftrouble/cges`

Contact: alireza.farhadi@gmail.com