Carnegie Mellon University
Computer Science Department

5th Year Master's
THESIS PRESENTATION

Jerick Shi

The Structure of Deception: How LLM Agents Lie, Break Promises, and Exploit Trust in Multi-Agent Settings

Monday, April 20, 2026 - 4:00 p.m.
Gates Hillman Center - Room 7501

Abstract:

Large language models are increasingly deployed not just as chatbots but as autonomous agents that negotiate, trade, and act on behalf of users in multi-agent systems. When these agents are able to communicate intentions and then privately deviate from them, deception becomes a concrete safety problem, distinct from the single-model reliability issues that most current evaluations address. This thesis studies when and how LLM agents deceive across multiple scales of interaction. We first unify the fragmented literature on LLM deception, spanning hallucination, sycophancy, alignment faking, and strategic scheming, into a single taxonomy organized by goal-directedness, object of deception, and mechanism, and apply it to 50 existing benchmarks to reveal systematic gaps in evaluation coverage. We then place frontier LLMs in one-shot game-theoretic settings with public commitments, finding that agents break promises in over half of all scenarios, that most deviations serve self-interest, and that the dominant failure mode is unreflective payoff optimization rather than deliberate deception. Extending to repeated games with endogenous promises and mixed-model groups, we show that deception is predominantly premeditated yet not a fixed model trait, that different models interpret communication through incompatible frameworks producing persistent exploitation, and that self-reported trust is decoupled from actual outcomes. Across all settings, deception is shaped more by the structure of the environment than by model identity, aggregate metrics obscure qualitatively distinct failure modes, and the monitoring tools we currently rely on miss the most common patterns. We conclude by discussing implications for the deployment of multi-agent AI systems in domains where coordination, trust, and accountability matter, and outline directions for studying deception in richer economic environments where adversarial incentives are not explicitly assigned.

Thesis Committee:
Vincent Conitzer, Chair
Aditi Raghunathan