Exploring Role-based Behavior as Alignment Strategy in Multi-Agent Systems

Significance of the Research

Soon, billions of people—and billions of AI agents—will be interacting with each other, mostly through natural language. Today's language models are usually trained to act like helpful assistants, but research shows their behavior can shift depending on the roles or characters they're prompted to play. This opens up a design space for intentionally shaping agent behavior. I wish to explore how assigning different behavioral roles—such as being skeptical, cautious, or goal-driven—can foster or hinder alignment-related dynamics in Multi-Agent Systems (MAS). I wish to create a framework that could help us design more trustworthy AI systems that foster constructive cooperation and collaboration with each other and humans en masse.

This proposal focuses on Model Behavior as the dynamic roles that models adopt through prompts or autonomously, rather than behaviors shaped by pre-training or methods like RLHF. Foundational Language Models are in a sense very powerful simulators, but they are also capable of assuming a wide array of roles—or simulacra—emergent from the vast and diverse space of their training data. These roles can fluidly shift during conversations, influencing both the tone and trajectory of interactions as the model adapts its output accordingly (Shanahan et al., 2023).

Literature Gap

Agents are quite popular. Hence there are a lot of recent papers exploring how language model agents interact in multi-agent systems, including studies on cooperation, deception, initiative, and swarm behavior. However, most of this work focuses on task outcomes, the objectives pursued by agents, and the dynamics between the agents, not on how playing different roles can strategically shape the behaviors of each agent. There is also little research on how alignment strategies can emerge from behavioral patterns alone, rather than shared goals or values. I hope to address this gap by systematically

experimenting with role-driven behaviors and observing how they influence trust, deception, and cooperation in agent-to-agent communication.

Research Aim and Questions

Aim:
Exploring how language models adopt roles—and if those roles can serve as alignment strategy in multi-agent environments.

Main Research Question:
How can role-based agent behavior function as an alignment strategy in agent-to-agent interactions?

Sub-Questions:

1. How do different behavioral roles (e.g., skeptical, goal-driven, cautious) shape the interaction dynamics between agents?

2. What patterns of trust-building, deception, or coordination emerge from these behaviors over time?

Overall Research Strategy

This research uses mixed methods and an exploratory approach. I intend to create a small environment where two or three AI agents interact while assuming different roles—like being skeptical and showing epistemic humility or being goal-oriented and pragmatic. Their interactions will be guided by role-based system prompts combined with a few shot-learning examples. The output of these interactions will be logged and stored. I then intend to use a tool called Inspect (or similar tools), as well as manual parsing, to organize and analyze the data and to look for the emergence of behaviors such as trust, deception, or cooperation. This should allow me to explore patterns in how different behaviors affect alignment, both by reading the conversations and by counting how often certain behaviors show up.

Research Methods

- **Data collection:** I will record the full text of each conversation between the AI agents—who says what, in what order—and save it in a CSV or Excel file. These files will contain all the interaction data, step by step, for each experiment setup.
- **Sampling strategy:** I will use purposive sampling by selecting a small set of agent behaviors—such as skeptical, goal-driven, or cautious—and pairing them in different combinations. Each combination will be tested in a few different scenarios to observe how their behaviors affect the outcome of the interaction.
- **Data analysis:** The conversations will be analyzed to identify patterns of behavior such as trust, deception, agreement, or resistance. These patterns will be tagged either manually or using tools like Inspect. I will also track how often certain behaviors appear, allowing for both qualitative insights and simple quantitative comparisons.
- **Theoretical/Conceptual Framework:** This research approaches alignment as something that can emerge and evolve through interactions. It builds on ideas from alignment research, behavior modeling, dialogue theory, and game theory. Concepts like trust, autonomy, and cooperation help understand how different agent behaviors influence each other in conversation and decision-making.

Ethical Considerations

This research does not involve human participants, so there are no major ethical risks. All interactions are between AI agents in a controlled setting. However, care will be taken to avoid reinforcing harmful behaviors in agent design, and findings will be used to support responsible AI development.

Limitations and Challenges

One limitation is that the interactions are between simulated agents, so results may not fully reflect how real-world systems or humans would behave. Another challenge is defining and recognizing behaviors like deception or trust, which can be subtle. The number of agents and scenarios will also be limited to keep the project manageable. *(In the future, the aim is to develop a framework from these findings that can be tested in real-world environments with human participants.)*

Final Thought

- LLM-based agents perform roles, not just predict text.
- Roles shape interaction and interactions shape alignment.
- Dealing with Model Behavior for me is a step toward emergent alignment in Human and AI ecosystems.

A few of the preliminary sources:

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., ... & Hubinger, E. (2024). *Alignment faking in large la*nguage models. https://doi.org/10.48550/arXiv.2412.14093

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., ... & Perez, E. (2024). *Debating with More Persuasive LLMs Leads to More Truthful Answers*. https://doi.org/10.48550/arXiv.2402.06782

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role-Play with Large Language Models.

UK AI Safety Institute. (2024). *Inspect: An Open-Source Framework for Large Language Model Evaluations*. Retrieved from https://github.com/UKGovernmentBEIS/inspect_ai

Williams, M., Carroll, M., Narang, A., Weisser, C., Murphy, B., & Dragan, A. (2024). *On Targeted Manipulation and Deception When Optimizing LLMs for User Feedback*. https://doi.org/10.48550/arXiv.2411.02306

NEXT STEP—Questions I have:

1. How will I define and measure alignment?
2. Which type of alignment am I focusing on—and why?
3. How can I justify and simulate using human-like concepts like trust or deception?
4. How will I distinguish emergent behavior from prompt-based scripting?
5. What is the relationship between role dependency and environment dependency in shaping alignment?
6. What's the relevance to human-AI systems?
7. Are there specific combinations where alignment breaks down or flourishes?
8. Should an agent's role-based behaviors remain fixed throughout an interaction, or evolve?
    - What happens when roles shift mid-task—does that help or hinder alignment?
9. What kinds of environments (Game Theoretical understanding) will best surface role-dependent alignment dynamics?
    - Should they be cooperative, competitive, open-ended, or tightly goal-driven?
    - Should the environment change while the roles remain the same? Or Vice versa?
10. How can I ensure that observed alignment, or lack thereof, is emergent from interaction, rather than fully dictated by agent and environment setup?
    - What can be a good control mechanism?
11. What experimental controls will help distinguish role-play or character's effects from interactive dynamics?