Characterizing Cognitive Dynamics in AI: Designing Virtual Worlds for a Comparative Analysis of Persona-Driven Behavior in Individual and Collaborative Agentic Systems

Sankaran, Sowmya (School: Albuquerque Academy)

As we move into an era where AI agents are prevalent in many aspects of our lives, from solving scientific problems to acting as therapists, the world faces an unquantified risk. Every day, over 8 million interactions occur between AI and users experiencing major life disruptions or acute mental health crises. Within this fragile ecosystem, 1.2 million users daily act based on distorted realities from AI interactions. This has led to consequences including suicide, hospitalization, psychosis, homicide, and deteriorating relationships. Current AI companies prioritize performance over safety, so agents are released and allowed to reshape human perception without us understanding their behavior. To address this, we focus on two main problems: (1) understanding how personality constraints affect decisions of individual AI agents, and (2) examining how personas can influence behavior and collaboration in team and social network-based scenarios. In this project, we designed a novel behavioral framework, first building virtual worlds for studying agents, systematically designing 100+ agents to create a diverse community. Then, we evaluated our agents across three levels- (1) analyzing how agentic persona constraints affect behavioral tendencies, (2) examining team behavior in collaborative scenarios, and (3) modeling the social networks of agentic systems. Our analysis showed consistency in emergent behavioral patterns across all three levels. Finally, we translated these behavioral insights into a mobile app which deploys agents that are verified to be helpful and supportive, allowing for reliable and trustworthy human-agent interaction. Understanding agentic behavior provides crucial insights, with implications for the design of socially robust, safe, and reliable AI agents.

Awards Won: