Adaptive Agents for Mixed-Initiative Human-AI Collaborations

Manisha Natarajan

Georgia Institute of Technology
Atlanta, GA, USA
manisha.natarajan@cc.gatech.edu

Abstract

Efficient human-agent collaboration requires understanding each other's capabilities and establishing appropriate reliance. My thesis focuses on optimizing performance in mixed-initiative settings, where humans and agents dynamically contribute to decisions and actions. I first explore key factors shaping human reliance on decision-support agents, then examine how agents can model this reliance to initiate actions. My proposed work aims to enable agents to jointly provide decision and action support in multi-objective tasks, using bi-directional communication to enhance collaboration.

Introduction

Artificial Intelligence (AI) agents can significantly improve human performance by offering suggestions or assisting with tasks across diverse domains—from everyday activities to critical applications such as healthcare and disaster response. However, both humans and agents face limitations due to incomplete knowledge or physical and computational constraints. Optimizing task performance requires flexible collaborations where humans and agents can dynamically contribute based on their capabilities. This approach, known as *mixed-initiative* interaction, offers a promising strategy to improve joint task performance (Jiang and Arkin 2015). The key challenge, however, is ensuring that agents can adaptively assume initiative without compromising user trust.

Trust, defined as an attitude that an agent will help achieve an individual's goals in uncertain situations (Lee and See 2004), governs how users rely on agents. Miscalibrated trust—due to misunderstandings about the agent or task—can lead to users over- or under-relying on agents. In mixed-initiative interactions, trust determines whether users comply with agent initiatives, and miscalibrated trust disrupts the effective use of human and agent strengths, reducing task performance and leading to potentially severe consequences (Dismukes, Berman, and Loukopoulos 2007).

My thesis addresses the challenges of mixed-initiative interactions by developing adaptive agents that model user reliance to enhance human-agent teamwork. First, I investigate how trust and reliance are shaped by various agent, user, and task attributes when users interact with decision-support agents that offer recommendations. Next, I design adaptive agents that actively assist users by assessing reliance to determine when to take initiative. Building on these insights, my proposed work aims to create agents that provide suggestions and active assistance in complex, multi-objective environments, using bi-directional communication to foster appropriate reliance. By modeling user reliance and strategically leveraging human and AI capabilities, my research offers valuable insights for designing AI systems to effectively support users in complex scenarios.

User Reliance on Decision-Support Agents

I first examine how users depend on Intelligent Decision-Support (IDS) agents, which offer recommendations while users retain full task control (Natarajan and Gombolay 2020). Although IDS agents are intended to enhance decision-making, users tend to over-rely on agents, even when the recommendations are incorrect. Conversely, under-reliance can occur when users distrust agent capabilities, resulting in missed opportunities for improved decision outcomes (Lee and See 2004). Therefore, identifying factors that mitigate such inappropriate reliance is essential for optimizing human-agent team performance.

In my prior work, I conducted two studies with 195 subjects (N=75 in-person and N=120 online) to characterize users' inappropriate reliance on suboptimal IDS agents (Natarajan and Gombolay 2020). My work was the first to jointly investigate how various agent attributes (e.g., embodiment, feedback style, timing of suggestions), task attributes (e.g., outcome delays), and user attributes (e.g., task proficiency) affect trust and reliance. I used an interactive math quiz and the Concentration card game to represent non-sequential and sequential decision-making tasks for the study. I measured trust with a validated scale (Jian, Bisantz, and Drury 2000) and inappropriate reliance by tracking when users accepted poor or rejected correct suggestions.

The studies showed that the perceived anthropomorphism (human-likeness) and the agent's feedback after decision-support failure (e.g., the agent apologizes or holds users accountable) significantly influenced trust and reliance. Moreover, delaying the agent's suggestions reduced over-reliance, even when users could not immediately verify their decisions. These findings show that adapting agent behavior can effectively mitigate inappropriate reliance.

Active Assistance in Mixed-Initiative Teams

After examining how users rely on IDS agents that offer suggestions, I expanded my research to explore how users depend on agents that actively contribute to task outcomes through direct interventions and explanations (Natarajan et al. 2024; Natarajan, Xue, and Gombolay 2023). Inspired by search-and-rescue missions, we designed a navigation task where humans and agents have limited visibility and thus perform suboptimally. The user controls the agent toward a goal, but the agent can intervene by taking control or pausing to find safer or shorter paths. The agent may also use pre-defined explanations to provide context, and users can accept or oppose the agent's interventions.

Our approach, Bayes-POMCP, builds on prior work (Katt, Oliehoek, and Amato 2017) and is designed to optimize collaboration with users without requiring prior interaction data. Bayes-POMCP models user reliance and leverages Monte Carlo tree search (MCTS) to determine when and how the agent should intervene based on the current task context. Our approach continuously refines the user reliance model through approximate Bayesian updates, allowing the agent to adapt its intervention strategy in real-time.

We conducted two user studies to evaluate the effectiveness of agent interventions in mixed-initiative teaming. The first study (N=30) tested different intervention strategies (stop, take-control, with or without explanations) and found that agent interventions significantly improved team performance, especially when the agent took control. However, user preferences for intervention strategies varied, highlighting the need for adaptive intervention. The second study (with N=28 new users) evaluated the adaptive capabilities of Bayes-POMCP and found that Bayes-POMCP significantly improved team performance and user perceptions (trust, agent likeability) compared to baseline methods. These results show that online adaptation, by estimating user reliance and considering human-agent capability asymmetries can greatly enhance mixed-initiative collaborations when humans and agents have incomplete knowledge.

Bi-directional Communication in Multi-Objective Settings

My proposed work aims to design agents that coordinate with humans in multi-objective settings. Studying multi-objective settings is crucial, as several real-world tasks often require agents to manage concurrent or competing objectives. For instance, in cooking, agents must balance preparing multiple dishes at once while minimizing resource waste. Bi-directional communication is essential in such scenarios to avoid misunderstandings between humans and agents and foster appropriate reliance. Building on insights from my prior work (Sections 1 and 2), I plan to develop agents capable of both offering suggestions and actively collaborating with users in complex, multi-objective settings.

My work will focus on implementing bi-directional communication using language—the most intuitive form of interaction for users. To achieve this, agents must interpret and generate meaningful language. I will integrate pre-defined concept grounding (mapping language to task components) into the Bayes-POMCP framework, allowing agents to reason about communication and actions. However, reasoning in multi-objective settings increases the problem search space. To address the added complexity, I will develop novel MCTS heuristics that enable agents to balance task progress while quickly learning user preferences to assist efficiently.

I will explore scenarios such as collaborative cooking and search-and-rescue, where task success depends on completing multiple objectives. These settings require dynamic control sharing, with humans directing agents or allowing autonomous operation while agents intervene as needed. My work will be the first to explore how agents can assist through communication and collaboration in multi-objective, mixed-initiative settings.

To validate my approach, I will conduct simulation-based pilot studies and physical robot experiments to demonstrate scalability and real-time collaboration. Currently, I am testing different search heuristics and plan to complete simulation experiments before the AAAI Doctoral Consortium. In the future, I also aim to integrate foundation models for more adaptive natural language communication and examine the ethical implications of agent interventions in safety-critical tasks. Ultimately, my work seeks to pave the way for more fluent human-AI collaboration in complex, multi-objective settings by leveraging the complementary strengths of humans and AI agents and fostering appropriate reliance through effective communication.

References

Dismukes, K.; Berman, B. A.; and Loukopoulos, L. D. 2007. *The limits of expertise: Rethinking pilot error and the causes of airline accidents*. Ashgate Publishing, Ltd.

Jian, J.-Y.; Bisantz, A.; and Drury, C. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*.

Jiang, S.; and Arkin, R. C. 2015. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International conference on systems, man, and cybernetics*, 954–961. IEEE.

Katt, S.; Oliehoek, F. A.; and Amato, C. 2017. Learning in POMDPs with Monte Carlo tree search. In *International Conference on Machine Learning*, 1819–1827. PMLR.

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1): 50–80. PMID: 15151155.

Natarajan, M.; and Gombolay, M. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 33–42.

Natarajan, M.; Xue, C.; and Gombolay, M. 2023. Mixed-Initative Human-Robot Teaming under Suboptimality. In *2023 AAAI Fall Symposium Series*.

Natarajan, M.; Xue, C.; van Waveren, S.; Feigh, K.; and Gombolay, M. 2024. Mixed-Initiative Human-Robot Teaming under Suboptimality with Online Bayesian Adaptation. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 1454–1462.