huaol@mit.edu

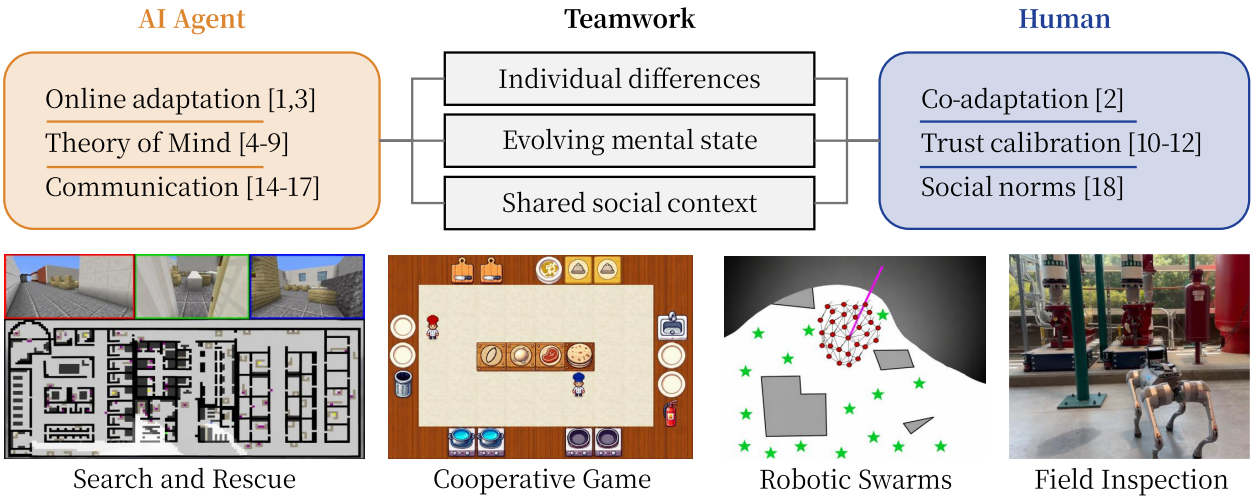
Huao Li

www.huao-li.com

Research Statement

As AI agents take on increasingly complex tasks alongside people, they must evolve from passive tools into proactive team members. In **human-agent teamwork** (HAT), humans and agents perform interdependent tasks toward common goals, requiring agents to coordinate their decisions with those of their teammates. Yet many AI training pipelines treat humans as a homogeneous, stationary, and context-free source of supervision, while evaluation often focuses on isolated one-to-one interactions. Consequently, agents that perform well on average may fail particular user populations, misread changing beliefs and intentions, or behave inappropriately in collective settings. My research addresses this gap by incorporating **computational models** of human into agent learning and decision-making, enabling agents to **adapt** their behavior for effective teamwork.

I pursue this agenda through three levels of modeling. Models of **individual differences** capture persistent traits and preferences, allowing agents to adapt to particular partners [1–3]. Models of **mental states** capture what changes during interaction, requiring agents to infer their partners' beliefs and intentions from limited observations and update their behavior accordingly [4–8]. **Social context** provides the common ground needed for communication and the norms that guide appropriate behavior [16–18]. Across these levels, my work follows a common computational recipe: represent relevant human factors quantitatively, infer them from interaction data or prior knowledge, learn agent policies conditioned on those models, and finally evaluate agents with humans in collaborative tasks.



Learning to collaborate with diverse humans

An agent policy that works well with one person may not work well with another. I build agents that recognize how a particular human tends to act and adapt accordingly. My earlier work inferred a human's strategy online and selected a complementary reinforcement learning policy from a pre-trained library [1, 2]. In human-agent experiments, this approach improved team performance by 42.3% over baselines. Because a predefined library may not cover actual human behavior, our subsequent NeurIPS work replaced it with a latent strategy space learned by a generative model from diverse agent populations, enabling adaptation to previously unseen partners [3]. Individual differences are signal, not noise around an average: strategy types learned from a population let an agent recognize partners it has never met, whereas an agent trained against the average is optimized for a user who does not exist.

Adaptation also does not flow in one direction. Participants paired with an adapting agent revised their own strategies more than those paired with a fixed-policy agent, and this **co-adaptation** improved joint outcomes [2]. The agent's behavior therefore changes the evidence from which it learns about its partner. Capturing this reciprocal dynamic requires modeling how people interpret and respond to the agent during interaction.

Inferring the evolving mental states of teammates

Beyond stable traits, people's beliefs, intentions, and trust change during an interaction, and none of these **mental states** is directly observable. In successful human teams, members reason about teammates' perspective and share information needed for coordination without overwhelming one another. I developed computational Theory of Mind (ToM) models combining Bayesian methods and neural networks to sequentially infer teammates' beliefs and intentions from their actions and verbal communication [4–6]. These models allow an AI advisor to intervene when team members' mental models diverge. In the DARPA ASIST program, we evaluated this approach with human teams performing simulated search and rescue missions. Low-competency teams supported by our advisor achieved 13.8% greater performance gains across missions than those supported by a human advisor [7]. In my later EMNLP work, I applied the similar idea of explicit belief representation to teams of multiple large language model (LLM) agents, and observed improved social reasoning and collaboration capabilities in embodied tasks [8].

Human reliance on artificial teammates poses a complementary inference problem: people must also form accurate expectations of the agent. **Trust calibration** becomes especially challenging when the reasoning behind an agent's behavior is opaque. In human supervision of robot swarms, I found that trust and intervention tracked observable collective dynamics such as heading variation and spatial dispersion, which task performance alone did not explain [10, 11]. To calibrate trust, I model it as a latent dynamical state estimated with a personalized Kalman filter that captures the lag between perceived evidence and trust updates [12], and automatically change the supervisory control mode based on estimated user trust [11]. Calibration can also run in the other direction: rather than adjusting to a person's trust, an agent can make itself easier to understand. Our work on explainable reinforcement learning, which received the Best Paper Award at AIES [13], showed that explaining agent behavior helps align user expectations with agent capability.

Grounding interaction in shared social context

Understanding a partner also requires accounting for the social context in which the interaction unfolds. Human teams rarely build common ground from scratch. They inherit a shared language that allows members to exchange information without continually negotiating its meaning. Shared norms likewise shape expectations about appropriate behavior. Agents trained in isolation do not learn such knowledge by default. My work uses LLMs as a source of **social priors** and investigates how these priors can guide interaction while remaining open to correction.

Multi-agent reinforcement learning (MARL) agents can develop efficient but idiosyncratic communication protocols that humans cannot interpret [14, 15]. Task reward alone does not constrain messages to carry meanings understood by human partners. I therefore make human **interpretability** an explicit learning target by aligning agents' messages with natural-language messages produced by LLM agents in the same situations [16]. This approach combines environmental reinforcement with semantic grounding, balancing task utility and generalization to unseen teammates and task situations.

Communication also requires robots to interpret ambiguous human instructions, which often leave references vague and commonsense, safety, and social constraints unstated. LLMs tend to fill these gaps with overconfident guesses. I developed a Bayesian active learning framework that treats LLM interpretations as initial hypotheses and maintains a probabilistic model over possible user intents and select clarification questions that maximize expected **uncertainty reduction** [17]. The inferred intent guides a formal planner to generate robot trajectories satisfying the clarified task specification. Our approach improves task satisfaction with fewer clarification rounds than LLM-based baselines, and we deployed the system on a quadruped robot for inspection tasks at a GE Vernova factory. This work connects human-robot communication with verifiable planning and was selected for a spotlight presentation at CoRL 2026.

Correctly understanding human intent does not by itself establish appropriate agent behavior, which also depends on social context. In a survey of culturally diverse users about domestic robots [18], I found systematic differences in normative expectations across social contexts and cultural backgrounds. Alignment methods that treat aggregated preference labels as a uniform standard of "human values" risk reproducing the biases of their training populations. These findings motivate **pluralistic value alignment**, in which agents account for whose expectations guide behavior and how those expectations depend on context.

Future Research Directions

My work provides computational approaches to adapting across people, inferring mental states, and grounding interaction in social context. The next challenge is to integrate these insights into agents that learn continually through real-world deployment, sustaining effective collaboration as their partners and social environments evolve.

Evaluating social intelligence through interaction. As AI systems grow more general and opaque, their behavior must be studied empirically [19], yet most benchmarks remain static datasets. My work on LLM ToM [8] was among the first to evaluate social reasoning through dynamic multi-agent collaboration rather than static text vignettes. I will extend this approach to multi-human, multi-agent systems, where errors propagate through communication and individually capable agents can reach unreliable collective decisions. Psychometric theory provides a foundation for defining which social reasoning capability an interaction measures and testing whether that measurement generalizes across contexts. Because optimizing agents for user satisfaction alone can encourage sycophancy, I will also measure whether agents improve team performance, whether users' reliance matches agent capability, and whether users retain agency over their decisions in high-stakes domains such as education, health, and manufacturing.

User modeling for proactive assistance and mutual adaptation. Proactive assistance requires an agent to anticipate what a user needs, but those needs can shift in response to the agent itself. My co-adaptation studies [2] establish the importance of this reciprocal influence, while learned latent strategies represent differences in coordination behavior [3]. Building on this foundation, I will develop hierarchical user models that distinguish stable tendencies from current mental states and interpret both within the surrounding social context. The central question is whether an agent can distinguish persistent preferences from behavioral changes caused by its own actions. I will study how agents and humans update their models of one another, and anticipate the consequences of action over time. Because learned strategy spaces remain task-specific, I will also test whether independently measured traits predict coordination styles across tasks, providing agents with a prior over unfamiliar partners. This extends personalization toward sustained mutual adaptation, where effective assistance depends on how both partners learn from their interaction.

Social learning for self-improvement and value alignment. As agents increasingly learn from their own experience, alignment must account for the social feedback encountered during interaction. Human norms persist partly because others enforce them through their responses to behavior. I will investigate how agents can learn appropriate policies from such norm-enforcing behaviors, extending my work on ToM-based social influence as a learning signal for prosocial behavior [9]. However, social pressure can also enforce conformity or encourage collusion, especially when participants have conflicting interests. My research will therefore extend beyond collaboration to examine when emergent conventions support human interests and when they undermine them. Building on my postdoctoral work, I will explore how neuro-symbolic methods connect socially learned policies to explicit, verifiable constraints where requirements can be specified [17]. Governance must also address whose expectations shape learning, through empirical evaluation and mechanisms that let affected people assess and influence the resulting behavior according to their own values and needs.

Long-term Vision and Collaboration

Across my work, the methods have changed but the core question has not: how should an agent account for the people and society around it? My long-term goal is to redefine how AI agents are trained and evaluated, extending beyond static datasets and isolated interactions toward collective learning embedded in society. Building this requires a science of human-agent teamwork that no single field can supply, which motivates my interest in collaborating across departments.

This agenda supports foundational and applied research. Computational models of human-agent teamwork align with NSF's Human-Centered Computing program, while autonomy, trust, and robotics connect to DoD programs, extending my contributions to DARPA-, ARL-, and AFOSR-supported research. Industrial partnerships offer opportunities in human-robot collaboration with companies such as Honda and GE Vernova, alongside interpretability, alignment, and socially responsible AI in the technology sector.

References

[1] **Huao Li**, Dana Hughes, Michael Lewis, and Katia Sycara. Individual adaptation in teamwork. In *Annual Meeting of the Cognitive Science Society*, volume 42, 2020. Abstract.

[2] **Huao Li**, Tianwei Ni, Siddharth Agrawal, Fan Jia, Suhas Raja, Yikang Gui, Dana Hughes, Michael Lewis, and Katia Sycara. Individualized mutual adaptation in human-agent teams. *IEEE Transactions on Human-Machine Systems*, 51(6):706–714, 2021.

[3] Benjamin Li, Shuyang Shi, Lucia Romero, **Huao Li**, Yaqi Xie, Woojun Kim, Stefanos Nikolaidis, Charles Lewis, Katia Sycara, and Simon Stepputtis. Adaptively coordinating with novel partners via learned latent strategies. *Conference on Neural Information Processing Systems*, 38:1268–1295, 2025.

[4] **Huao Li**, Long Le, Max Chis, Keyang Zheng, Dana Hughes, Michael Lewis, and Katia Sycara. Sequential theory of mind modeling in team search and rescue tasks. In *AAAI Fall Symposium on Computational Theory of Mind for Human-Machine Teams*, pages 158–172. Springer, 2021.

[5] **Huao Li**, Ini Oguntola, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind modeling in search and rescue teams. In *IEEE International Conference on Robot and Human Interactive Communication*, pages 483–489, 2022.

[6] **Huao Li**, Yao Fan, Keyang Zheng, Michael Lewis, and Katia Sycara. Personalized decision supports based on theory of mind modeling and explainable reinforcement learning. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 4865–4870, 2023.

[7] Dana Hughes, **Huao Li**, Max Chis, Ini Oguntola, Simon Stepputtis, Keyang Zheng, Joseph Campbell, Katia Sycara, and Michael Lewis. A framework for intervention based team support in time critical tasks. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2023.

[8] **Huao Li**, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 180–192, 2023.

[9] **Huao Li**, Michael Lewis, and Katia Sycara. Lg-tom: Language grounded theory of mind modeling for multi-agent communication. 2026. Under review.

[10] Changjoo Nam, **Huao Li**, Shen Li, Michael Lewis, and Katia Sycara. Trust of humans in supervisory control of swarm robots with varied levels of autonomy. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 825–830. IEEE, 2018.

[11] Changjoo Nam, Phillip Walker, **Huao Li**, Michael Lewis, and Katia Sycara. Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human-Machine Systems*, 50(3):194–204, 2020.

[12] **Huao Li**, Michael Lewis, and Katia Sycara. A kalman estimation model of human trust in supervisory control of robotic swarms. In *Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 329–333. SAGE Publications Sage CA: Los Angeles, CA, 2020.

[13] Rahul Iyer, Yuezhang Li, **Huao Li**, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018.

[14] Mycal Tucker, **Huao Li**, Siddharth Agrawal, Dana Hughes, Katia Sycara, Michael Lewis, and Julie A. Shah. Emergent discrete communication in semantic spaces. In *Conference on Neural Information Processing Systems*, 2021.

[15] Seth Karten, Mycal Tucker, **Huao Li**, Siva Kailas, Michael Lewis, and Katia Sycara. Interpretable learned emergent communication for human–agent teams. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1801–1811, 2023.

[16] **Huao Li**, Hossein Nourkhiz Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi-Pari, Michael Lewis, and Katia P. Sycara. Language grounded multi-agent reinforcement learning with human-interpretable communication. In *Conference on Neural Information Processing Systems*, 2024.

[17] **Huao Li**, Carson Sobolewski, Augustinos D Saravanos, Tan Yew Teck, John Karigiannis, and Chuchu Fan. Bayesian active learning for intent disambiguation in interactive robot planning. In *Conference on Robot Learning*, 2026.

[18] **Huao Li**, Stephanie Milani, Vigneshram Krishnamoorthy, Michael Lewis, and Katia Sycara. Perceptions of domestic robots' normative behavior across cultures. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 345–351, 2019.

[19] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.