Research Statement

Tianmin Shu

No other species possesses a social intelligence quite like that of humans. Our ability to understand one another's minds and actions, and to interact with one another in rich and complex ways, is the basis for much of our success, from governments to symphonies to the scientific enterprise. My research goal is to advance **human-centered AI** by engineering **machine social intelligence** to build socially intelligent systems that can understand, reason about, and interact with humans in real-world settings.

Social intelligence is a highly interdisciplinary subject. To build AI systems that can achieve human-level social intelligence, we need insights and techniques from different fields. My interdisciplinary training in computer science, statistics, and cognitive science allows me to connect computer vision, machine learning, robotics, and social cognition to study machine social intelligence. In particular, I take inspiration from social cognition to identify the developmental roadmap of social intelligence and introduce benchmarks that systematically evaluate social intelligence in machines. I then combine deep learning, planning, reinforcement learning, and probabilistic inference to create cognitively inspired machine learning and AI approaches for two key building blocks of social intelligence: i) social scene understanding and ii) multi-agent cooperation. I have demonstrated that these approaches are resilient and data efficient, generalize well in unseen situations, and can be deployed to real-world systems (e.g., drones and humanoid robots).

I will summarize my work for the two key building blocks of machine social intelligence and discuss directions for future research.

1 Social Scene Understanding

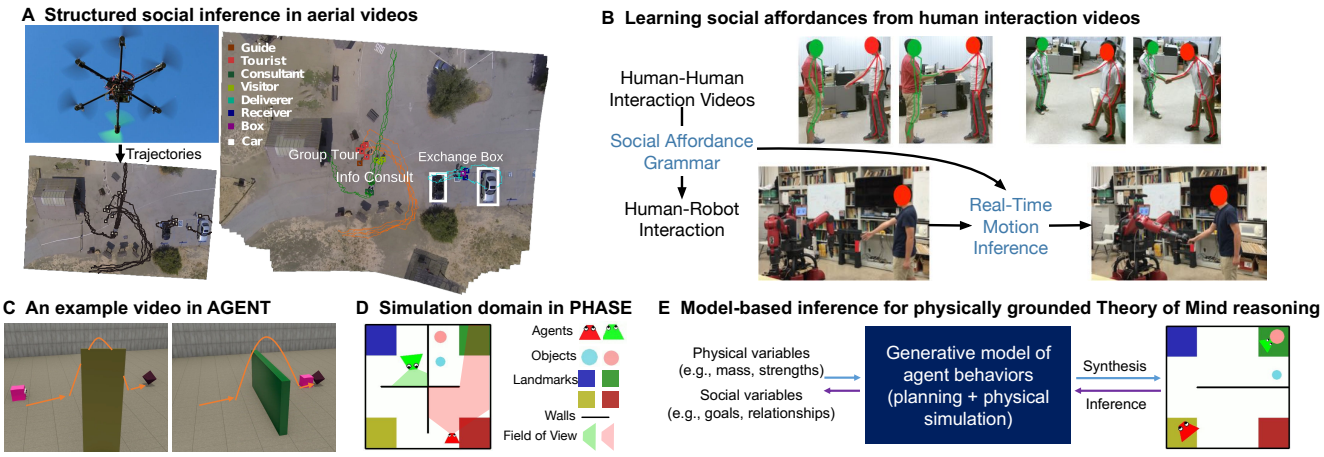

Figure 1: (**A**) I devised a grammar model to jointly infer of social groups, events, and roles from aerial videos captured by a drone [19]. (**B**) A similar grammar model can also learn social affordances from human interaction videos to enable human-robot social interactions. (**C, D**) Diagnostic benchmarks, AGENT [11] and PHASE [7], for physically grounded Theory of Mind reasoning. To understand the videos in these benchmarks, we need to combine Theory of Mind with intuitive physics. For example, in (**C**), we can imagine what is behind the occluder (right) after watching the agent's actions (left) since we understand the physical condition in which a rational agent should jump; and in (**D**), we need to understand how an agent acts with respect to the objects and obstacles in the environment as well as other agents. (**E**) Model-based inference for engineering human-level physically grounded Theory of Mind reasoning [13, 7, 11].

Human visual perception includes the ability to understand not only the physical environment but also the people in it. By observing other people's behaviors, we can easily understand their mental states (e.g., goals, beliefs, and desires) as well as their social relationships with one another. To reverse

engineer humans' abilities to understand one another, I work on social scene understanding, where I build cognitively inspired models that can make sense of human activities in videos by inferring the underlying social structures and reasoning about individuals' mental states. I have also collaborated with cognitive scientists and developmental psychologists at MIT and Harvard University to create diagnostic machine benchmarks following the experimental designs used in human social perception studies.

Recognizing social groups, events, and roles from real-world videos. Humans can represent social scenes in rich and complex ways—we can detect social groups, recognize the social events that groups engage in, and assign roles to individuals in those groups. In [19, 14, 15], I developed an approach to jointly infer social groups, events, and roles from real-world videos. I proposed a visually grounded spatiotemporal grammar as a structured representation of social interactions, which can be learned from a small amount of training data. The learned grammar model combines bottom-up proposals (e.g., sub-event detection) and top-down dependency parsing based on the spatial relationships between individuals and the temporal relationships between sub-events. By combining these two processes, the learned grammar model can holistically parse a video of human activities in a hierarchy, from motions to roles to events and finally to groups. To evaluate this approach, I created an aerial video dataset consisting of complex real-world social interactions in an open, public area with detailed ground-truth annotations (Figure 1**A**). The experimental results show that our approach significantly outperforms bottom-up–only baselines that directly map visual features to social judgments and is robust to noisy trajectories extracted from the aerial videos. In [18], I extended this approach by proposing structured social activity recognition enabled by energy-based modeling and a graph neural network, which boosts the performance of the previous grammar-based model by incorporating deeply learned spatiotemporal representations. This demonstrates the power of structured social scene understanding, which can take advantage of both structured reasoning and deep representation learning.

Learning social affordances from videos. In [16, 12], I devised an algorithm to learn social affordances (appropriate actions given a social context) in the form of spatiotemporal grammar from RGB-D videos of real-life human interactions (Figure 1**B**). The proposed weakly supervised grammar learning can automatically construct a hierarchical representation of a human-human interaction with the long-term joint sub-tasks of both agents and the short-term atomic actions of individual agents. The learned grammar allows us to transfer knowledge about social etiquette to human-aware robot motion planning, enabling a robot to engage in human-robot social interactions, such as shaking hands or handing over an object.

Physically grounded Theory of Mind reasoning. Social scene understanding goes beyond detecting and recognizing social events. In [13, 7, 11], I studied physically grounded Theory of Mind (ToM) reasoning, that is, inferring the mental states of agents from their complex social interactions under physical dynamics. I hypothesized that such reasoning can be built upon the understanding of how rational agents plan their actions to achieve goals with respect to physical dynamics and constraints (i.e., intuitive psychology), as well as the core knowledge of objects and physics (i.e., intuitive physics). I devised a model-based inference approach, termed Generative Social-Physical Inference (GSPI), which jointly infers the goals, relationships, and strengths of agents using a hierarchical planner and a physics engine as the generative model of agents and objects (Figure 1**E**). Taking inspiration from classic experiments designed for human social perception [6, 4], I also proposed two diagnostic benchmarks, AGENT (Action, Goal, Efficiency, coNstraint, uTility) and PHASE (PHysically-grounded Abstract Social Events), as shown in Figure 1**CD**. Each benchmark consists of a large-scale dataset of animations, depicting agents moving under various physical conditions, interacting with objects and with one another. With these datasets, we designed tasks in which models need to infer the mental states of agents and predict future trajectories. Our experiments show that our model-based inference significantly outperformed existing state-of-the-art methods due to its ability to generalize to unseen social and physical scenarios and estimate the uncertainty in inference.

Figure 2: (**A**) The VRKitchen platform allows human users to use VR devices to interact with objects and agents in simulated kitchens [1]. (**B**) We developed another platform, VirtualHome-Social, to simulate multi-agent household activities in realistic virtual apartments [9]. Based on this platform, we also proposed a new embodied AI assistance challenge, Watch-And-Help [9, 10]. To solve this challenge, I devised a novel online assistance approach, which combines neural networks, planning, and probabilistic inference [10]. (**C, top**) AI assistants in the real world must understand user preferences. For instance, to help a user to set up a desk, an AI agent needs to infer how this user wants to set up a desk (i.e., the goal specification behind this task). (**C, bottom**) To accurately infer such goal specifications, the AI agent should be able to actively communicate with users and utilize user feedback to efficiently update its understanding of the goal [8].

2 Multi-agent Cooperation

Our social interactions are guided by how we perceive one another. We help other people when we recognize their goals and the difficulties they might have in reaching them; we further communicate with them if we are uncertain about their true intentions. Therefore, I also study how we may use social scene understanding to guide multi-agent cooperation.

Embodied human-AI cooperation. In my research on embodied human-AI cooperation, I aim to engineer socially intelligent agents that can infer humans' mental states and collaboratively plan to work with humans in complex settings that are beyond traditional in-lab environments. To achieve this goal, we developed realistic multi-agent virtual platforms—VRKitchen [1] and VirtualHome-Social [9]—to collect datasets of human activities through virtual reality and online crowdsourcing and to train and test embodied agents along with simulated humans or real humans (Figure 2**AB**). Unlike traditional in-lab settings, such platforms allow us to create a large set of diverse environments that are close to the real world. Besides human-AI cooperation, our platforms have been used in studies in other areas, such as computer vision, smart homes, and robot planning.

In addition to building virtual platforms, we introduced a new embodied AI assistance challenge, Watch-And-Help [9, 10], in which embodied assistants must simultaneously watch humans' actions, infer humans' goals, and assist humans in reaching their goals (Figure 2**B**). This is inspired by how young children can help others by inferring others' intentions [22]. To solve this challenge, I proposed a novel online assistance approach: Neurally-guided Online Probabilistic Assistance (NOPA). NOPA consists of two main components: neurally guided online goal inference and an uncertainty-aware helping planner. For online goal inference, we trained a goal proposal network to produce possible goal hypotheses and evaluate the proposals using model-based inference. We then devised a hierarchical planning algorithm that can identify useful subgoals that may be necessary for reaching a range of possible goals. The helper agent enabled by NOPA can robustly update its inference and adapt its helping plans to the changing level of uncertainty in real time. The success of NOPA demonstrates the value of neural-symbolic models for building socially intelligent AI assistants in complex settings, in which neural networks can offer fast inference speed, and symbolic reasoning can ensure the robustness of inference and planning.

Active user preference learning. For AI agents to assist humans in the real world, they must have the ability to adapt to any user's preferences. For instance, to help a user organize a desk, an AI agent needs to understand how this particular user wants to set up a desk (Figure 2**C**). To this end, my

recent work [8] studied how AI agents can discover the goal specification for any task that a user may want to perform and reach the same goal in new environments. We formulated this problem as reward learning, in which the agent needs to watch a single demonstration for a task and learn a reward function to describe the goal of this task. This is a challenging learning problem since it is often unclear what the goal specification is from a single demonstration. Inspired by cognitive science studies on how children learn by forming and testing hypotheses [5], we designed an active reward learning algorithm that can propose different hypotheses about what the goal specification is and generate informative queries for the human user to verify the hypotheses. We conducted a user study to learn the spatial goals of object rearrangement tasks (a type of common robotics task) via graph-based reward functions. The results suggest that after only a small number of queries with a human user, our approach can drastically improve the reward function initially learned from a single demonstration. The final reward function allows an AI agent to perform the same task efficiently in unseen test environments.

Mind-aware reinforcement learning for ad hoc teaming. In [17], I explored the possibility of learning implicit Theory of Mind reasoning for reinforcement learning (RL) to achieve ad hoc teaming. Combining RL and self-supervised learning, I trained a belief tracker for an RL agent to learn to infer implicit representations of other agents' mental states and skills based on past observations accumulated during interactions with them. I found that RL agents jointly trained with this belief tracker could successfully adapt to ad hoc teammates on the fly and achieve a better generalization to unseen combinations and numbers of agents compared to common RL baselines. This suggests that Theory of Mind capacity may emerge via learning through experiences.

3 Future Directions

My current work has focused on the most fundamental aspects of social reasoning, such as inferring goals, desires, and beliefs. However, humans can make far more **complex and adaptable social inferences** than the current models. For instance, we can infer how other people interpret our own beliefs (i.e., nested belief inference) and judge whether we have mutual understanding with one another. We can also learn to adapt to different cultures and social norms. These abilities are not only crucial for understanding human interactions and communication but also fundamental for building trustworthy AI systems. I have taken initial steps in this direction, such as modeling nested inference in a new multi-agent framework, social MDPs [20, 21]. In the future, I intend to study i) how to create efficient inference algorithms that can conduct nested belief inference rapidly in complex social interactions and ii) how to develop machine learning methods to acquire knowledge about norms and conventions from little data.

Another key direction that I intend to explore more in the future is developing general frameworks for **embodied social intelligence**, including cooperation, communication, and social learning in realistic environments, by combining vision, language, and robotics. These frameworks will include multi-modal interactions between AI agents as well as between humans and AI agents. For this, we need new formalism and algorithms for multi-agent interaction and learning. We also need to build better embodied AI platforms that have not only high-fidelity physical simulation but also realistic human behavior modeling.

To achieve successful human-AI cooperation, both the human user and the AI agent need to understand each other. This requires an **alignment between humans and AI** to reach a mutual understanding of each other's mental models. In my past work, I have attempted to solve this problem through communication. In [3], we proposed a new type of motion planning method to generate expressive robot motion as a type of nonverbal communication to calibrate humans' beliefs about a robot's true physical capabilities. In [2], we introduced a novel explainable AI framework for generating multimodal messages to communicate with humans after detecting discrepancies between humans' beliefs and agents' own beliefs [2]. In both cases, humans found the AI partners more helpful and trustworthy when communication was enabled by our mutual mental modeling. For future work, I will investigate more general algorithmic alignment methodologies.

Finally, children develop increasingly sophisticated social intelligence through diverse experiences. Taking inspiration from studies on the origins and development of human social intelligence, I plan to study the role of learning in social intelligence. I am particularly interested in building **continuous machine learners** that can develop stronger social intelligence from human-like experiences accumulated through interactions with the physical world and other agents, using only a minimum set of inductive biases, such as core knowledge of intuitive physics and intuitive psychology.

In sum, engineering human-level machine social intelligence is crucial for building AI systems that can understand humans and interact with them in safe and productive ways. Successful deployment of such socially intelligent systems in real-world applications (such as autonomous vehicles, service robots, virtual companions in AR/VR, and AI teachers) will have a tremendous impact on our daily lives and on our society. To realize this future, I believe that we as AI researchers must pay more attention to the social aspect of AI and invest more efforts into the study of machine social intelligence. I also believe that we must bring together ideas and advances from different fields (including AI, robotics, cognitive science, and developmental psychology) through interdisciplinary collaboration to further the research on machine social intelligence.

References

[1] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, and S.-C. Zhu. VRKitchen: An interactive 3D environment for learning real life cooking tasks. In *ICML Workshop on Reinforcement Learning for Real Life (RL4RealLife)*, 2019.

[2] X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu. Joint mind modeling for explanation generation in complex human-robot collaborative tasks. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020.

[3] X. Gao, L. Yuan, T. Shu, H. Lu, and S.-C. Zhu. Show me what you can do: Capability calibration on reachable workspace for human-robot collaboration. *Under Review*, 2021.

[4] G. Gergely and G. Csibra. Teleological reasoning in infancy: The naïve theory of rational action. *Trends Cogn. Sci.*, 7(7):287–292, 2003.

[5] A. Gopnik and H. M. Wellman. Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085, 2012.

[6] F. Heider and M. Simmel. An experimental study of apparent behavior. *American Journal of Psychology*, 57(2):243–259, 1944.

[7] A. Netanyahu, T. Shu, B. Katz, A. Barbu, and J. B. Tenenbaum. PHASE: Physically-grounded abstract social events for machine social perception. In *35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[8] A. Netanyahu, T. Shu, J. Tenenbaum, and P. Agrawal. Discovering generalizable spatial goal representations via graph-based active reward learning. In *International Conference on Machine Learning*, 2022.

[9] X. Puig, T. Shu, S. Li, Z. Wang, J. B. Tenenbaum, S. Fidler, and A. Torralba. Watch-And-Help: A challenge for social perception and human-AI collaboration. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[10] X. Puig, T. Shu, J. B. Tenenbaum, and A. Torralba. NOPA: A challenge for social perception and human-ai collaboration. In *Under review*, 2022.

[11] T. Shu, A. Bhandwaldar, C. Gan, K. A. Smith, S. Liu, D. Gutfreund, E. Spelke, J. B. Tenenbaum, and T. D. Ullman. AGENT: A benchmark for core psychological reasoning. In *The 38th International Conference on Machine Learning (ICML)*, 2021.

[12] T. Shu, X. Gao, M. S. Ryoo, and S.-C. Zhu. Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[13] T. Shu, M. Kryven, T. D. Ullman, and J. B. Tenenbaum. Adventures in flatland: Perceiving social interactions under physical dynamics. In *42nd Annual Meeting of the Cognitive Science Society (CogSci)*, 2020.

[14] T. Shu, Y. Peng, L. Fan, H. Lu, and S.-C. Zhu. Inferring human interaction from motion trajectories in aerial videos. In *39th Annual Meeting of the Cognitive Science Society (CogSci)*, 2017.

[15] T. Shu, Y. Peng, L. Fan, H. Lu, and S.-C. Zhu. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in Cognitive Science*, 10(1):225–241, 2018.

[16] T. Shu, M. S. Ryoo, and S.-C. Zhu. Learning social affordance for human-robot interaction. In *The 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[17] T. Shu and Y. Tian. M^{3}RL: Mind-aware multi-agent management reinforcement learning. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[18] T. Shu, S. Todorovic, and S.-C. Zhu. CERN: Confidence-energy recurrent network for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu. Joint inference of groups, events and human roles in aerial videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] R. Tejwani, Y.-L. Kuo, T. Shu, B. Katz, and A. Barbu. Social interactions as recursive mdps. In *Under Review*, 2021.

[21] R. Tejwani, Y.-L. Kuo, T. Shu, B. Stankovits, D. Gutfreund, J. B. Tenenbaum, B. Katz, and A. Barbu. Incorporating rich social interactions into mdps. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[22] F. Warneken and M. Tomasello. Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765):1301–1303, 2006.