I have always been passionate about research and technical advances with positive societal impact, especially in the context of robotics. As robots transition from research labs to the real world, it is critical that we develop methods for establishing **mutual understanding between humans and robots** to bridge the communication, transparency, and collaboration gap between them. On one end, robots must be human-aware – comprehend human behavior, intentions, and expectations to act efficiently and safely in human environments. On the other end, their decisions and operational principles need to be transparent and explainable to humans, thereby fostering trust and facilitating effective collaboration. These challenges have motivated my general research interest in **human-robot interaction**, with a current focus on the **intersection of robotics and natural language processing**. My previous research revolved around enabling humans and robots to communicate and understand each other's intentions.

Leveraging Foundation Models for Behavior Explanation – *How humans understand robots*

To ensure humans understand the behaviors of intelligent agents like robots, these agents need to be able to explain the reasoning behind their decisions. Given human affinity for language communication, it would be beneficial if such explanations could be exchanged in natural language to enhance understanding and clarity. During my internship at Carnegie Mellon University, I worked with **Professor Katia Sycara** on generating linguistic explanations for black box agent policies. By leveraging the few-shot learning abilities of LLMs, I managed to develop a framework (Fig. 1) that generates natural language explanations for an agent's behavior based only on observations of states and actions, which also enables beneficial interactions such as clarification and counterfactual queries.

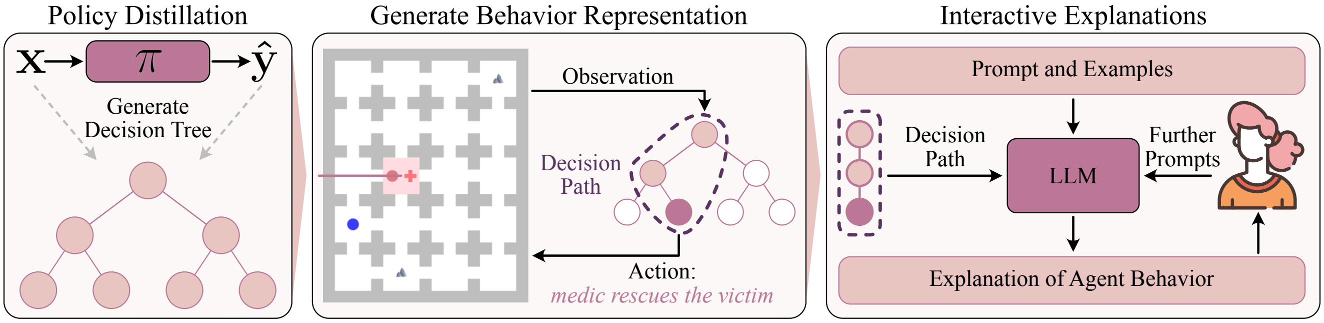

Figure 1: Overview of the three-step pipeline I established: 1) left: a black-box policy is distilled into a decision tree; 2) middle: a decision path is extracted from the tree for a given state which contains a set of decision rules used to derive the associated action; 3) right: an easily understandable natural language explanation is produced with LLM given the decision path. Lastly, a user can ask further clarification questions interactively.

Through distilling an agent's policy into a set of interpretable decision rules, I built a behavior representation of the policy that can be directly injected into a text prompt and reasoned with. I conducted empirical experiments and designed user studies to verify that 1) our approach generates explanations as helpful as those produced by a human domain expert. 2) our approach allows follow-up interactions that are helpful in understanding an agent's behavior. 3) our approach produces notably fewer hallucinations – an ever-present problem in LLMs where false information is presented as fact. This effort [1] has been presented at the **HmRI workshop at IROS 2023**.

Building on this state of the work, I further evaluated the performance of our approach by testing different policies and various categories of goal states collected through policy rollouts. To provide quantitative analysis beyond the user study, I measured the accuracies of the explanations in terms of policy-agnostic metrics and reaffirmed that our method consistently produces the most accurate explanations compared with other alternatives. To show that our framework is actually able to reason over and explain agent behaviors, I utilized the framework to predict next actions through follow-up queries and found that our method produces the best performance compared with alternative baselines. This work [2] has been submitted as a full conference paper to **IJCAI 2024 (Under Review)**.

Theory-of-Mind and Belief Maintenance in Human Environments – *How robots understand humans*

Robots face challenges in comprehending human environments, particularly due to the perceptual disparities and dynamic interactions of humans with their surroundings. For robots to develop human-awareness, they need to acquire a theory-of-mind (the ability to ascribe mental states to themselves and humans) and maintain a consistent belief about the world. At the University of Michigan, I worked with **Professor Joyce Chai** on a dataset that presents the challenge of perspective-taking on a home-set robot, focusing on scenarios where humans alter the location and state of objects. I designed scripts and collected data on a real-world robot.

Meanwhile, I spent a year setting up the real physical robot - the Tactile Intelligent Autonomous Ground Robot (TIAGo). I configured it to execute basic movements like grasping, picking, and placing, while also enabling it to collect sensory data across visual, auditory, and motional dimensions. Separately, I developed programs to direct the end-effector toward regions marked with ArUco identifiers. I also facilitated the robot to execute motions based on keyboard inputs and to recognize and act upon spoken commands.

Machine Learning Aided Sensory Perception – *How robots learn to perceive the world*

Early in my undergrad, I was curious about the correlation between artificial intelligence and perception and started my first research project with **Professor Zhi Yang**, focusing on enhancing the performance of the electronic nose. The electronic nose is a device that analyzes the chemical makeup of a sample and determines its unique signature. In this work, I applied Convolutional Neural Network Modeling, Recurrent Neural Network Modeling, and classical machine learning models to analyze the sensor signals for identifying gas components and deducing gas concentrations. I investigated the optimization of the sensor array and quantified the overall performance under various array sizes. I also designed an optimization criterion that assesses the advantages of adjusting sensor quantities. This work [3] has been submitted to **Sensors and Actuators: B (Under Review)**, a leading journal in chemical sensors.

References

[1] **X. Zhang**, Y. Guo, S. Stepputtis, K. Sycara, and J. Campbell, “Explaining agent behavior with large language models,” *IROS Human Multi-Robot Interaction Workshop (HmRI)*, 2023.

[2] **X. Zhang**, Y. Guo, S. Stepputtis, K. Sycara, and J. Campbell, “Understanding your agent: Leveraging large language models for behavior explanation,” in *International Conference on Learning Representations (ICLR) (under review)*, 2024.

[3] **X. Zhang***, T. Wang*, W. Ni, W. Lv, Y. Zhang, M. Zeng, J. Yang, Y. Su, N. Hu, and Z. Yang, “Sensor array optimization for the electronic nose via different deep learning methods,” in *Sensors and Actuators: B (under review)*, 2024.