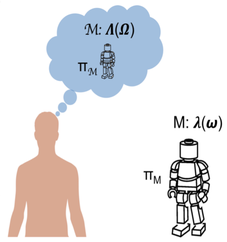
CAREER: When Reality Fails Expectations: Containing Reflective Domain Models for Human-Aware Planning and Learning of Robotic Teammates

Yu ("Tony") Zhang, Artzona State University

1. *Challenges & Motivation*

- Teammates have many conscious and subconscious expectations of others in terms of their plans or behaviors
- The expected domain model (a "reflection" for humans to generate expectations of the robot) and the true domain model may differ, leading to unmatched expectations, loss of situation awareness and trust

➢ This calls for generalized planning and learning methods for domain model reflective planning and learning



2. *Scientific Impact*

- A step towards ubiquitous collaborative robots with non-expert users, such as for autonomous cars, household robotic assistants, etc.
- Address the reflective model of robots to improve team situation awareness and trust; contributing to explainable AI and robotics

3. *Technical Impact*

- Generalize planning methods to real-world domains where the true domain model and the human's model of expectation are considered simultaneously
- Generalize learning methods that use human inputs to handle reflective model

4. *Technical Challenges*

- Model-reflective planning
  - *Explicable planning*
  - *Explanation generation*
  - *Robust planning under reflective model (the reflection is more accurate)*
- Model-reflective learning
  - *Reward learning with reflective model*
  - *Reinforcement learning with reflective model*

5. *Progress and Projections*

- Model-reflective planning
  - Active explicable planning (***IROS 2021***)
  - Safe and hierarchical explicable planning (under review)
  - Domain concretization (***RA-L 2022***)
- Model-reflective learning
  - Explicable policy search (***NeurIPS 2022***)
  - Learning from partial preferences for user-adaptive control (ongoing)
  - Preference-based learning with conflicting reflective models (ongoing)

5. *Broader Impact: Societal*

- *Ubiquitous collaborative robots require robotic technologies that support human-robot teaming*
- *Safety and trust issues*
- *Co-bot technology for improving our everyday life; public awareness*
- *Interpretable and explainable AI (AI explains complex behaviors and their rationale)*
- *Synergy with other programs*

6. *Broader Impact: Education*

- *Invited talk at IROS workshop*
- *Supervising graduate and undergraduate engineering projects (FURI & MORE) for students at ASU*
- *National Robotics Week 2022*
- *ASU Full circle article*
- *Planning a graduate and undergraduate class on "Human-Robot Interaction"*

7. *Broader Impact in numbers (1st & 2nd year)*

- *2 PhD students partially supported*
- *3 publications so far; more under review or in progress*
- *1 invited talk on the research topic: "Model-reflective planning and learning for human-robot teaming" (IROS 2022)*
- *1 public article on this research project*
- *1 Award committee*

2023 FRR & NRI Principal Investigators' Meeting
May 2-3, 2023

Award ID#: **2047186**