EAGER: Reconciling Model Discrepancies in Human-Robot Teams

Yu ("Tony") Zhang, Arizona State University

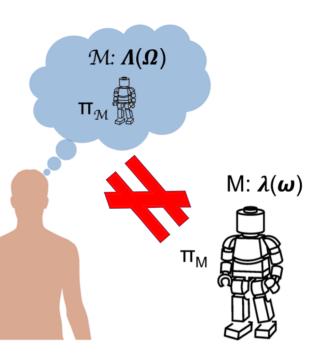
Award ID#: 1844524

Challenges and Motivation

- Teammates have many conscious and subconscious expectations of others in terms of their plans or behaviors
- The expected domain model for generating expectations and the true domain model may differ, leading to unmatched expectations, loss of situation awareness and trust

Technical Innovations

- Model reconciliation planning setting:
 - Explicable planning (implicit)
 - Explanation generation (explicit)



Discrepancies between the domain models can introduce a misalignment between the robot's behavior and its expectation

Scientific Impact

Generalize traditional planning methods for decision making under model discrepancies

Broader Impact

- Ubiquitous collaborative robots require robotic technologies that support human-robot teaming
- Synergies with NIH goals for robotics research
- Interpretable and explainable AI (AI explains complex behaviors and their rationale)

Education and Outreach

- Graduate class on "Human-aware robotics"
- Partially supported 3 PhD and 1 MS students, and projects for undergraduate students with awards
- Invited talks at industrial (e.g., Intel) and academia (e.g., IROS workshops) and various outreach activities (e.g., Intel ISEF, NRW)



Model Reconciliation Planning

Explicable Planning

Explicability metric

Explicability as Minimizing Distance from Expected Behavior, **AAMAS 2019**

 Approximate explicability measure using plan distance metrics for model reconciliation

Dynamic expectation

Generating Active Explicable Plans in Human-Robot Teaming, *IROS 2021*

• Bayesian framework for monitoring changing human expectation due to model updates and leveraging it for planning

Explanation Generation

Interpretation effort

Explain only plan segment

Online Explanation Generation for Planning tasks in Human-Robot Teaming, *IROS 2020*

• Breaks a plan explanation into multiple parts to explain at different time steps during the plan execution: interleaving explanation with plan execution

Optimize information order

Order Matters: Generating Progressive Explanations for Planning Tasks in Human-Robot Teaming, *ICRA 2021*

• Propose a learning method based on IRL to determine the preferred order of information in a plan explanation during explanation generation

Model Reconciliation Learning: What Is It You Really Want of Me? Generalized Reward Learning with Biased Beliefs about Domain Dynamics, AAAI 2020 Oral

Generalized reward learning with model discrepancies

Robot has an imperfect domain model: Domain concretization from examples: Addressing missing domain knowledge via robust planning, RA-L 2022

• Refine robot's model based on teacher's traces without knowing that it is not perfect