**Motivation & Contributions**

- For human-robot teaming scenarios, if the behaviors of robotic agents are incomprehensible to the humans, then it can impose cognitive load on humans and potentially introduce safety risks.
- In order to overcome these issues, in the plan synthesis process we not only consider planning model of the agent but also consider human’s interpretation of the robot’s behavior.
- This interpretation refers to human’s understanding of robot’s capabilities, mental states, etc.
- Differences between the actual robot model and human interpretation of the robot model can cause confusion and surprise when the human finds robot’s behavior different from his/her expectations.
- The difference in the model exists because human’s understanding about the robot’s model is often incomplete and inaccurate.
- A challenge in addressing this problem is that the human’s understanding of the agent’s model is inherently hidden and unknown.
- We propose a formulation to capture and learn this hidden model. We then integrate it in our planning process to generate plans as per human’s expectation of robot plans.

**Contributions:**
- Introduced the concept of explicability for robot task planning.
- Incorporated explicability as a heuristic in explicable plan generation process.
- Investigated two problem scenarios:
  - Human as a passive observer
  - Human-robot peer-to-peer teams.
- Evaluated the system for both scenarios with physical robot experiments.

**Problem Formulation (Human as Passive Observer)**

Given a goal, the objective is to find a robot plan that minimizes a weighted sum of cost of robot plan and differences between robot plan based on \( M_h \) and human’s expectation of robot plan based on \( M_h^* \).

\[
\begin{align*}
\text{arg min}_{\pi_{st}} & \quad \text{cost}(\pi_{st}, M_h) + \alpha \cdot \text{dist}(\pi_{st}, \pi_{st}^*) \\
\text{dist}(\pi_{st}, \pi_{st}^*) &= F(\pi_{st}) \quad \text{where} \quad F(\pi_{st}) = \text{C}(\pi_{st})\end{align*}
\]

\( L^* \) is the labeling scheme of the human for robot plans based on \( M_h^* \). \( L_h^* \) is the learned model of \( L^* \). We use linear chain conditional random fields as the graphical model for learning because of their abilities to model sequential data.

**Plan Synthesis using Explicability Heuristic**

**Experimental Analysis (Human as Passive Observer)**

- The robot’s goal is to build a tower of a certain height using blocks.
- There are two types of blocks, light and heavy, but that information is hidden from humans.
- Picking up the heavy blocks is costly than the light blocks for the robot.
- Hence, from the human’s perspective, the robot may sometimes choose seemingly more costly (i.e., longer) plans to build a tower.
- In this evaluation, we only use one task label “building tower”.
- For all testing problems, the labeling process results in 77.8% explicable actions for OPT and 97.3% explicable actions for FF-EXPD.
- The average explicability measures for FF-EXPD and OPT are 0.98 and 0.78, and the average scores are 9.65 and 6.92, respectively.

**Problem Formulation (Human as Active Collaborator)**

- The robot has access to it’s own planning model and approximate planning model of the human, \( M_h^* \).
- In the planning process, robot has to not only consider \( M_h^* \) but also the actual human planning model \( M_h \) which may be different from \( M_h^* \).
- Composite plan, \( \pi_c \), captures actions performed by both human and robot to achieve their goals.

\[
\begin{align*}
\arg\min_{\pi_{st}} & \quad \text{cost}(\pi_{st}, M_h) + \alpha \cdot \text{dist}(\pi_{st}, \pi_{st}^*) \\
\text{dist}(\pi_{st}, \pi_{st}^*) &= F(\pi_{st}) \quad \text{where} \quad F(\pi_{st}) = \text{C}(\pi_{st})\end{align*}
\]

\( L_c^* \) is the learned model of which takes labeled traces of composite plans as its training examples. Composite plans have alternate agent actions.

**Goal of robot is to form EAT and goal of human is to form PEN**