Teaching by Demonstration: Towards a Bayesian Approach to Explaining Robot Policies

Peizhu “Pam” Qian, Vaibhav Unhelkar
Department of Computer Science
Rice University
{pqian, unhelkar}@rice.edu
AI Safety Requires Transparency in AI Systems

- Robots interact with their end users to provide services in homes, schools, hospitals, and workplaces.
- To truly realize robots' benefits, ensure safety, and avoid unintended side-effects, it is essential that the end users understand the robots' behavior.

Teaching Humans about Robots

- Technical contribution: An algorithm (*Teacher*) to teach the *Robot* policy to a human user (*Student*).
- *Robot* behavior is summarized by its Markovian policy, which is unknown to the *Student*.
- *Student* knowledge is modeled as a probability distribution over possible *Robot* objectives.
- *Teacher* selects informational demonstrations to maximize the *Student’s* knowledge of the *Robot’s* policy, using a variant of Monte Carlo Tree Search.
Preliminary Results and Ongoing Steps

- Preliminary evaluations (in synthetic grid-world tasks) provide proof of concept for our approach.
- Encouraged by these results, the next steps of our ongoing research include applications to more complex tasks and evaluations with human users in the loop.

For more details: pqian@rice.edu