EAGER: Hierarchical Contrastive Explanations for Robot-Human Communication

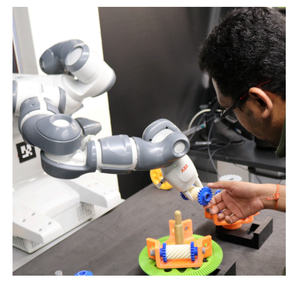
PI: Siddharth Srivastava, Arizona State University

Co-PI: Subbarao Kambhampati, Arizona State University

How would a non-AI/robotics expert determine what their robot can and can't do?

Understand what it's doing and why?

Reconfigure it for a desired objective?



Key Challenges

1. User needs to be able to ask the right questions to assess robot's capability for new tasks.
2. Robot needs to be able to *explain* itself. Explanations need to minimize the computational cost of processing information. This depends on the user's depth of knowledge.

Scientific Impact

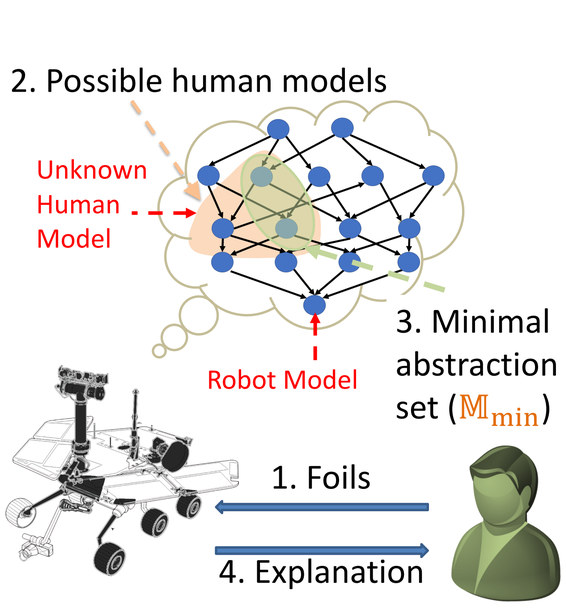
- Query-based model estimation can be used to compute understandable models for black-box, non-stationary cyber-physical systems.
- Theory of hierarchical abstractions for sequential decision-making can be used for formal verification of AI systems.
- Hierarchical explanations can be used for personalized training.

Solution Approach and Research Outcomes

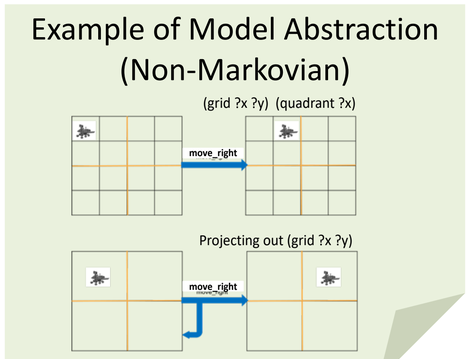
Use new theory of model abstractions to define lattice of abstract models [1].

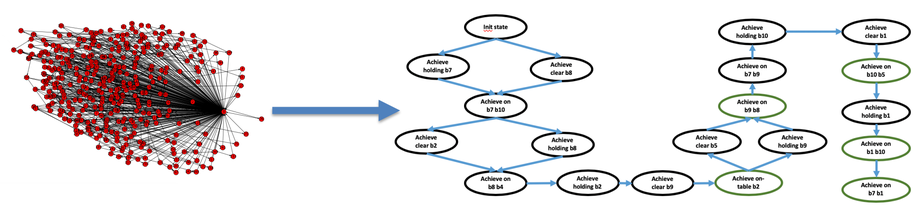
A. Use lattice properties to provide explanations that minimize user's computational cost of processing information [2, 3, 4]

B. Use lattice to compute hierarchical questioning strategy that constructs understandable model of robot's capabilities [5]

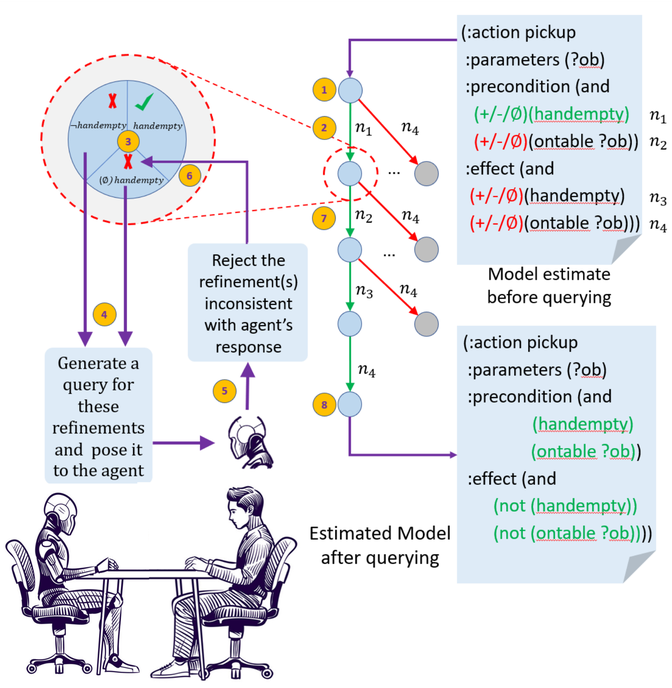


Computing personalized skill-aligned explanations of robot behavior [2,3]





Summarizing MDP policies using abstraction [5]

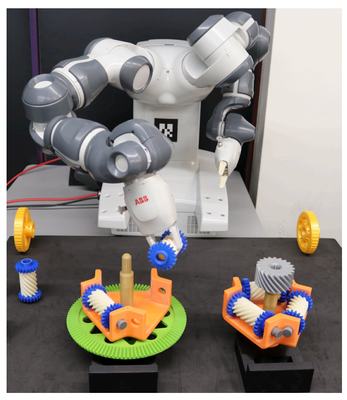


Enabling the user to ask the right questions [4] (black-box robot in non-stationary environment)

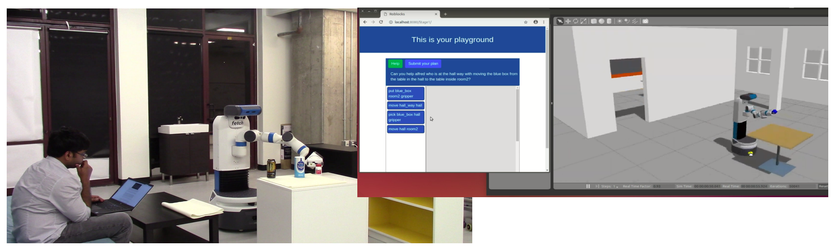
Broader Impact [Society]

Personalized, adaptive on- the-job training for working with AI robots

Increased worker employability and global competitiveness



Broader Impact [Education]



Adaptive AI/robotics education portals

Quantitative Impact

- ~2% (10%) of the US workforce has a bachelors' degree in computer science/mathematics (all science/engg)*
- This research develops foundations for enabling 90% of the workforce to use robots safely and effectively.

* NSF Science & Engineering Indicators 2018

1. *Metaphysics of Planning Domain Descriptions*. Siddharth Srivastava, Stuart Russell, Alessandro Pinto. In Proc. AAAI, 2016.
2. *Hierarchical Expertise Level Modeling for User-Specific Contrastive Explanations.* Sarath Sreedharan, Siddharth Srivastava, Subbarao Kambhampati. In Proc. IJCAI, 2018.
3. *Why Can't You Do That HAL? Explaining Unsolvability of Planning Tasks.* Sarath Sreedharan, Siddharth Srivastava, David Smith, Subbarao Kambhampati. In Proc. IJCAI, 2019.
4. *Learning Generalized Models by Interrogating Black-Box Autonomous Agents.* Pulkit Verma, Siddharth Srivastava. AAAI Workshop on Generalization in Planning, 2020.
5. *TLdR: Policy Summarization for Factored SSP Problems Using Temporal Abstractions.* Sarath Sreedharan, Siddharth Srivastava, Subbarao Kambhampati. In Proc. ICAPS, 2020 (to appear).
6. *Anytime Task and Motion Polices for Stochastic Environments*. Naman Shah, Kislay Kumar, Pranav Kamojhalla, Deepak Kala Vasudevan, Siddharth Srivastava. In Proc. ICRA 2020 (to appear).

2020 National Robotics Initiative (NRI) Principal Investigators' Meeting
FEBRUARY 27 - 28, 2020 | ARLINGTON, VIRGINIA

Award ID#: 1844325