EAGER: Hierarchical Contrastive Explanations for Robot-Human Communication

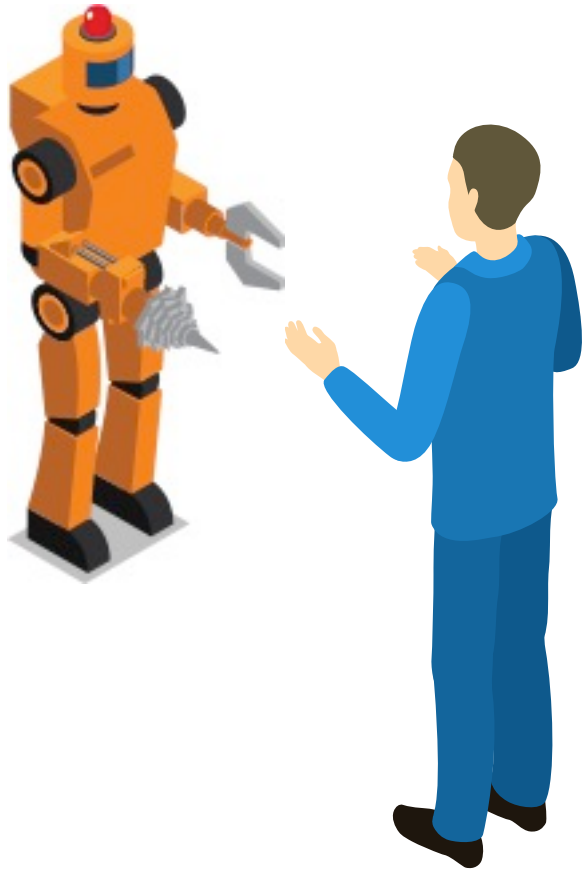
PI: Siddharth Srivastava

Context

Less than 2% of the workforce has college degrees in robotics/AI

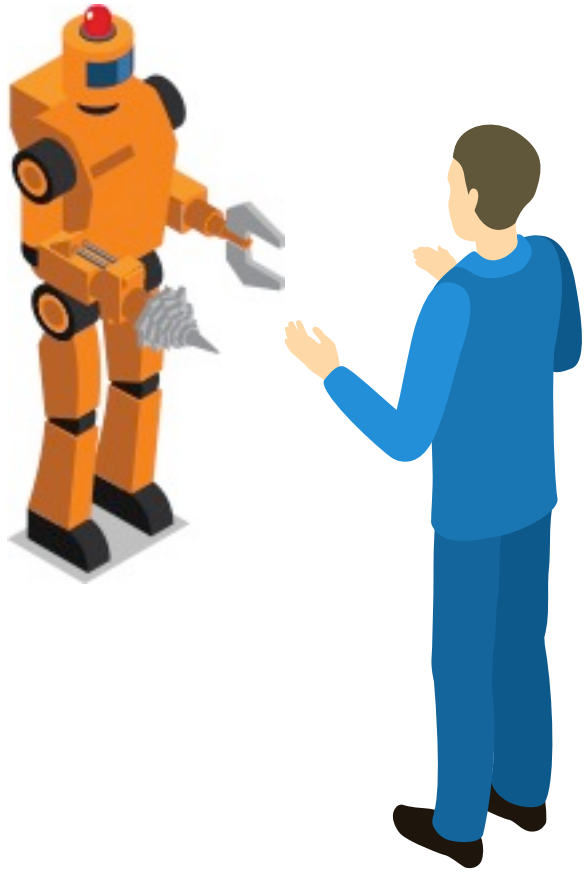
The Problem of Using Robots

- How would a non-AI/robotics expert use a robot?
  - Reconfigure the robot for new tasks
  - Understand robot's task-specific behavior
  - Estimate robot's capability for solving new tasks



The Problem of Using Robots

- How would a non-AI/robotics expert use a robot?
  - Reconfigure the robot for new tasks
  - Understand robot's task-specific behavior
  - Estimate robot's capability for solving new tasks



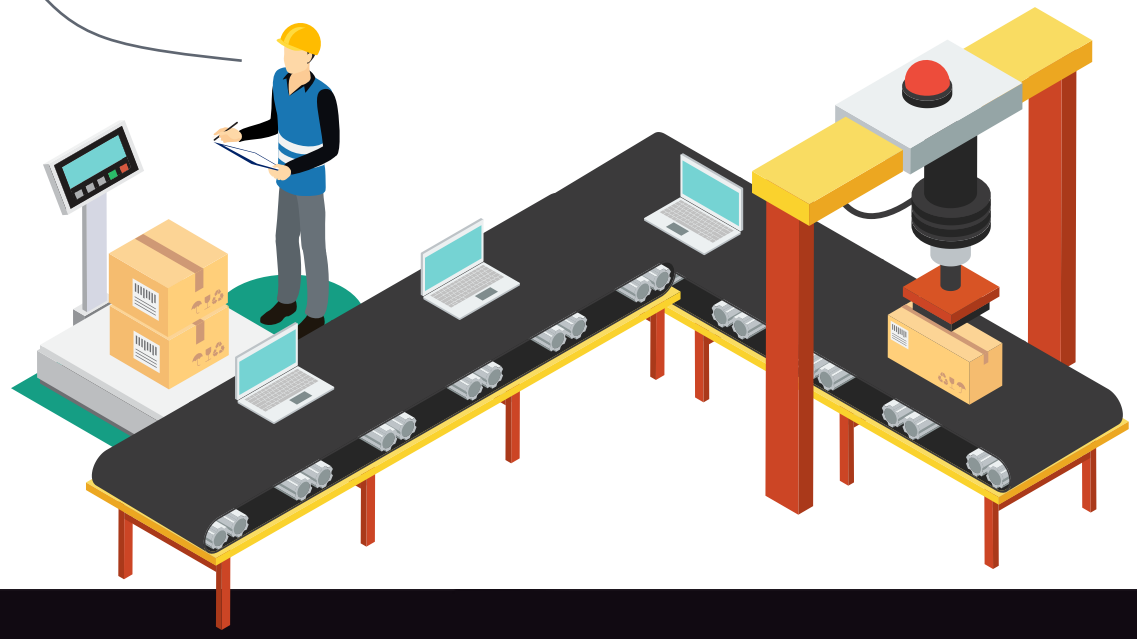
The Unexpected Solution

- Suppose the robot does something unexpected while carrying out a task
- User asks robot a question: Why are you doing X and not Y?

I'll change my gripper then bring you the package you asked.

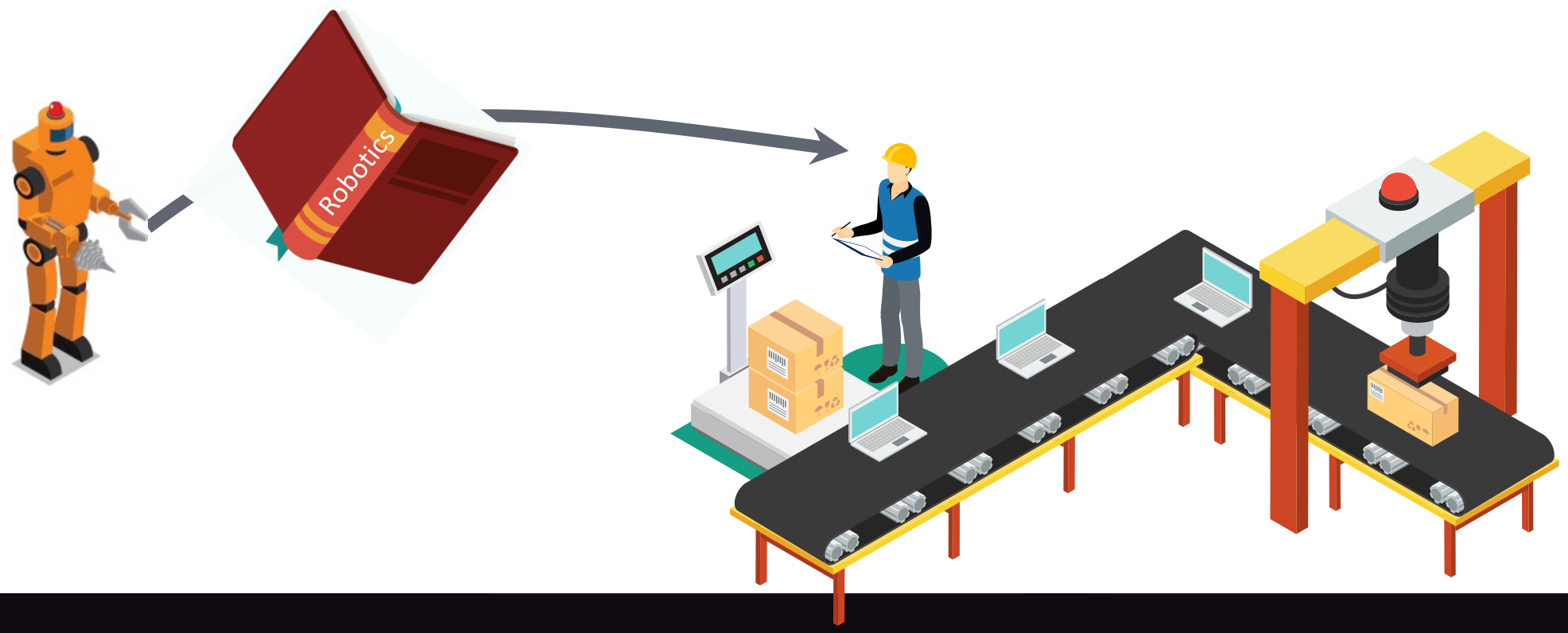
"Foil"

Why don't you bring me the package first?



The Unexpected Solution

- Suppose the robot does something unexpected while carrying out a task
- User asks robot a question: Why are you doing X and not Y?

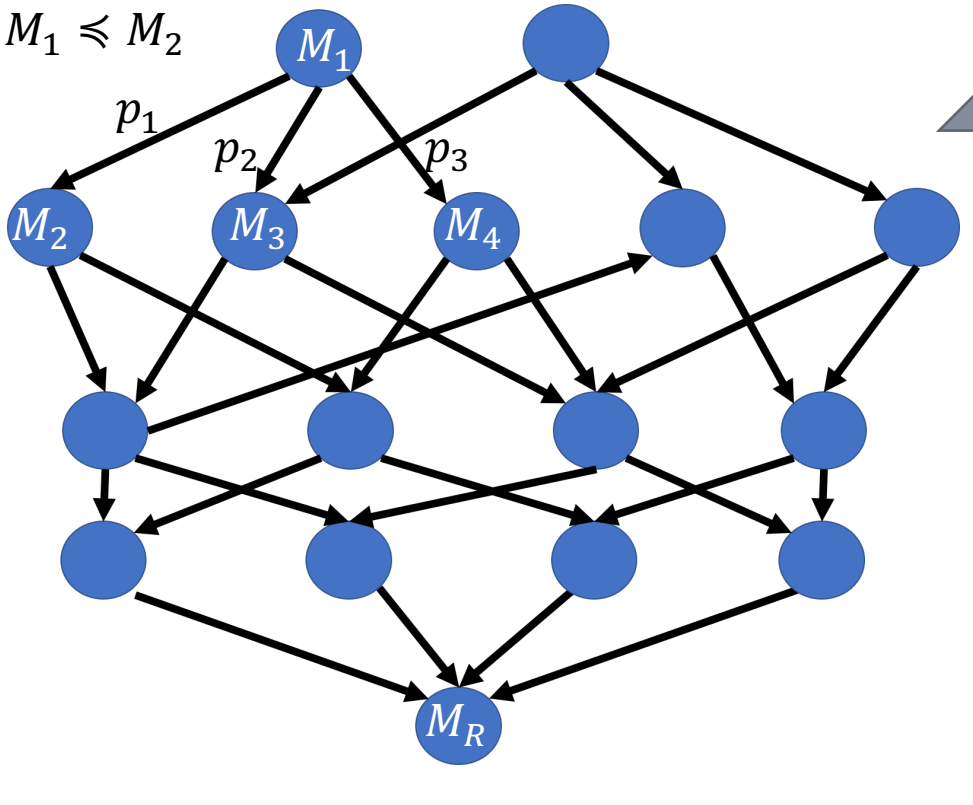


The Unexpected Solution

- Suppose the robot does something unexpected while carrying out a task
- User asks robot a question: Why are you doing X and not Y?
- Explanations go beyond providing information – need to be “easy to digest”

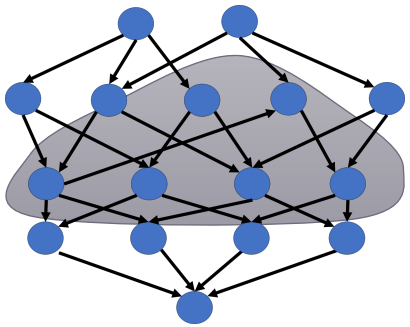
Hierarchical Abstractions for Explanations

- We treat explanation as the process of refining the user's state of knowledge
  - State of knowledge modeled using a lattice of abstract models



Hierarchical Abstractions for Explanations

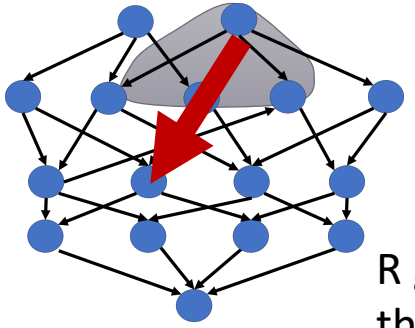
- We treat explanation as the process of refining the user's state of knowledge
- The user's model lies in a portion of the lattice where their foil holds true
- Explanation = minimal information to take H to a more precise model where the foil is inconsistent.
  - Use lattice properties to do this efficiently.



R's initial estimate of H's knowledge

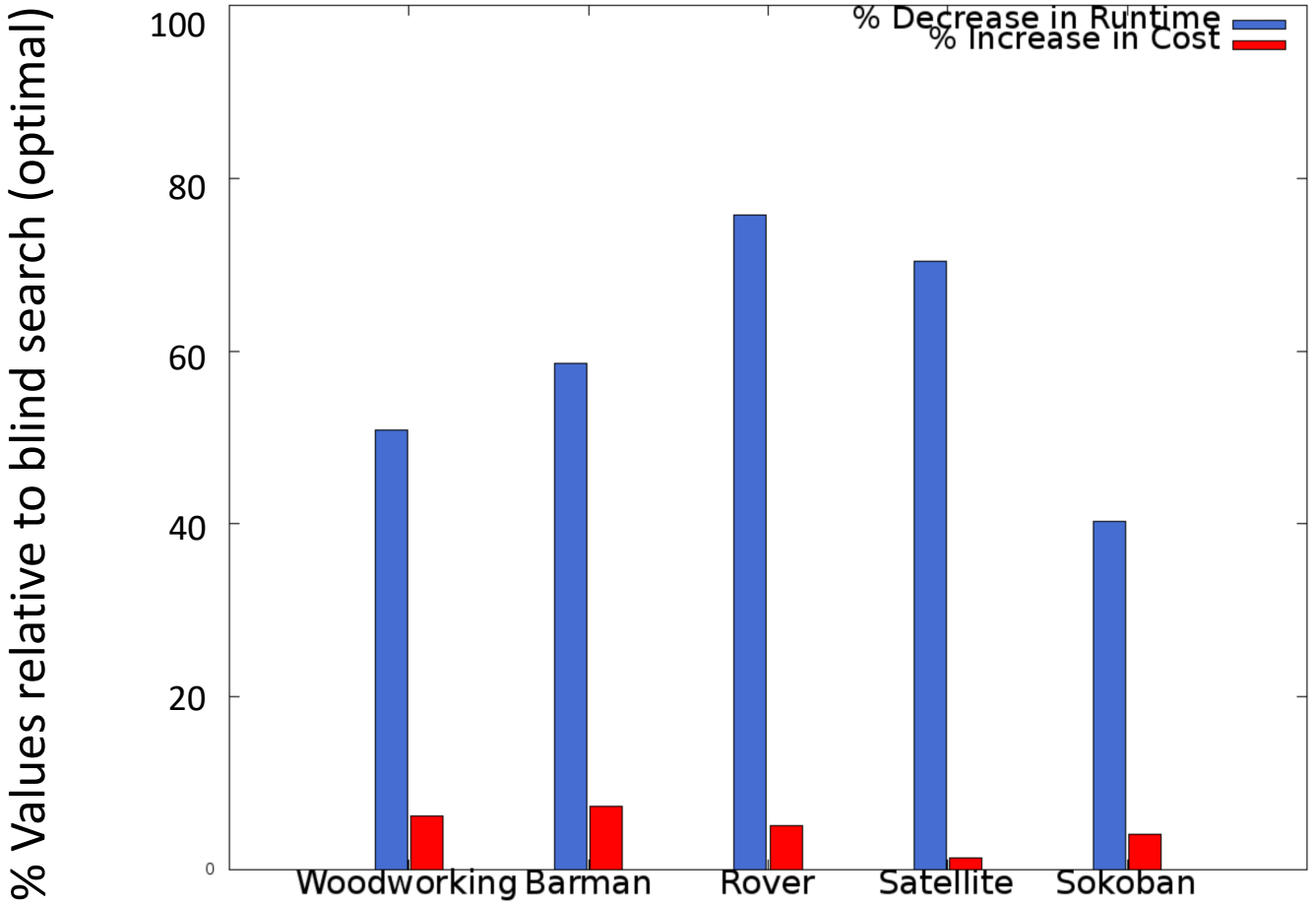
Why not do π?

π must be feasible according to H



R gives minimal information that makes foil infeasible

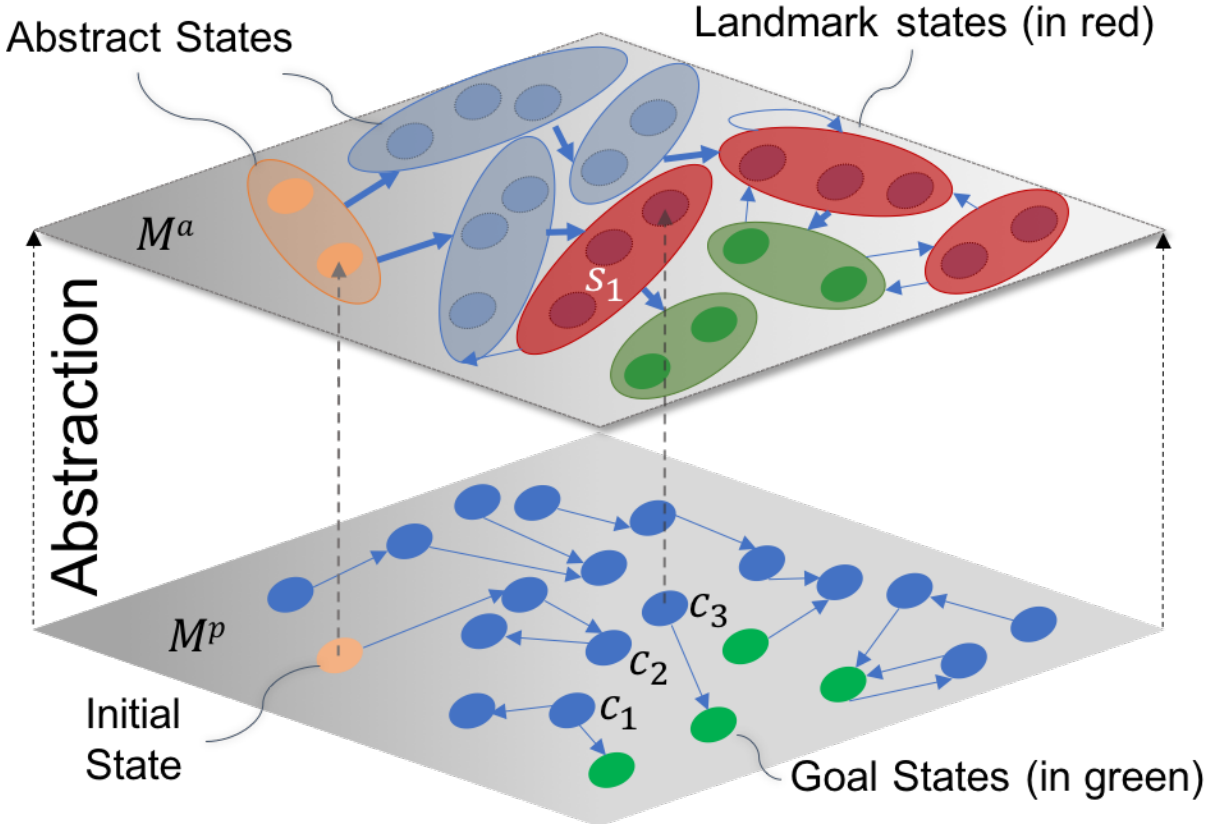
Greedy Algorithm: Performance Trade-off



[IJCAI 2018]

Hierarchical Abstractions for Explanations

- What if the robot finds the task unsolvable?
  - No behavior to explain!
  - Humans sometimes want to know "why couldn't you solve it?!"
- Same principles help robot explain unsolvability [IJCAI 2019]
- …and summarize MDP policies [ICAPS 2019]



Black-Box Robots, Non-Stationary Environments

What if we have a black-box robot?

Or the robot undergoes an “update”?

How would the user assess this robot’s capability?

Insight from human interactions: could one interview the robot?

Results on Black-Box Agents

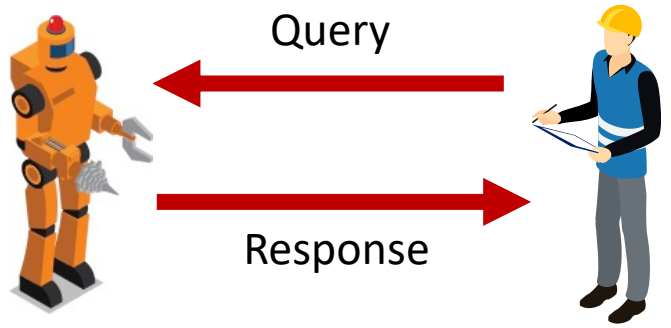
- A primitive query interface

  Query: What do you think will happen if you executed $\pi = \langle a_1, a_2, \ldots a_l \rangle$ in state s_i?

  Response: Execution will stop after k $(\leq l)$ steps in state s_f

  (supported by any simulator-based or analytical model-based robot)

- Key principle: use similar abstraction lattice to compute a top-down questioning strategy

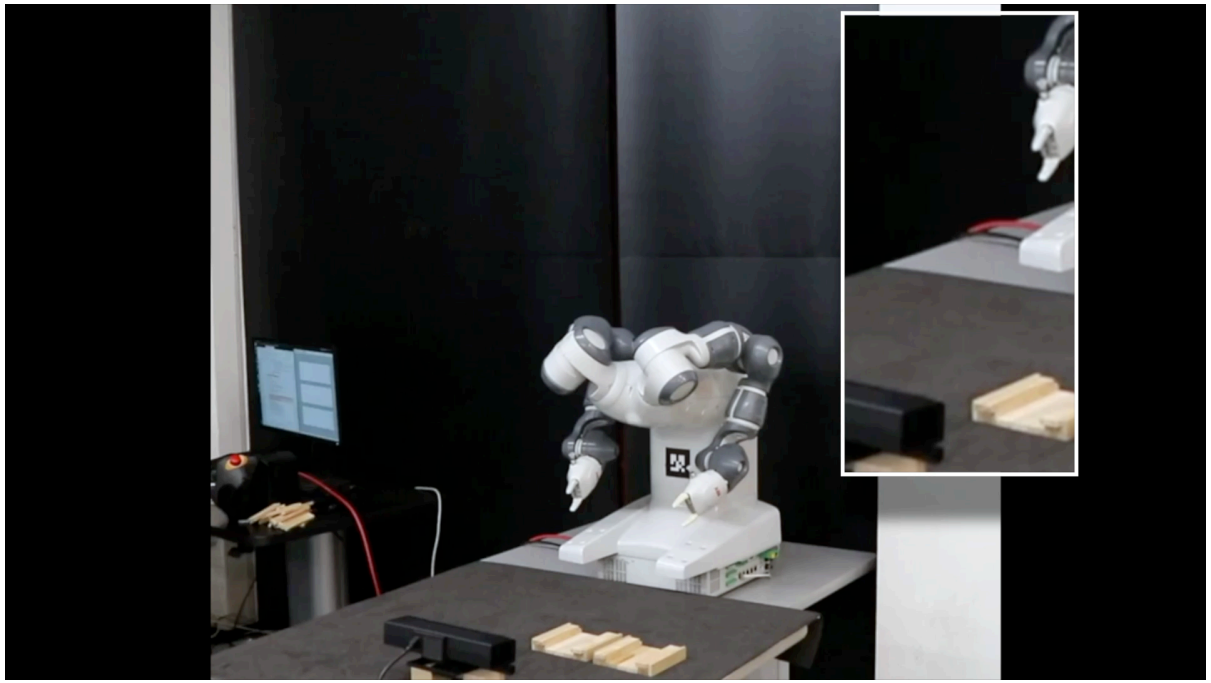


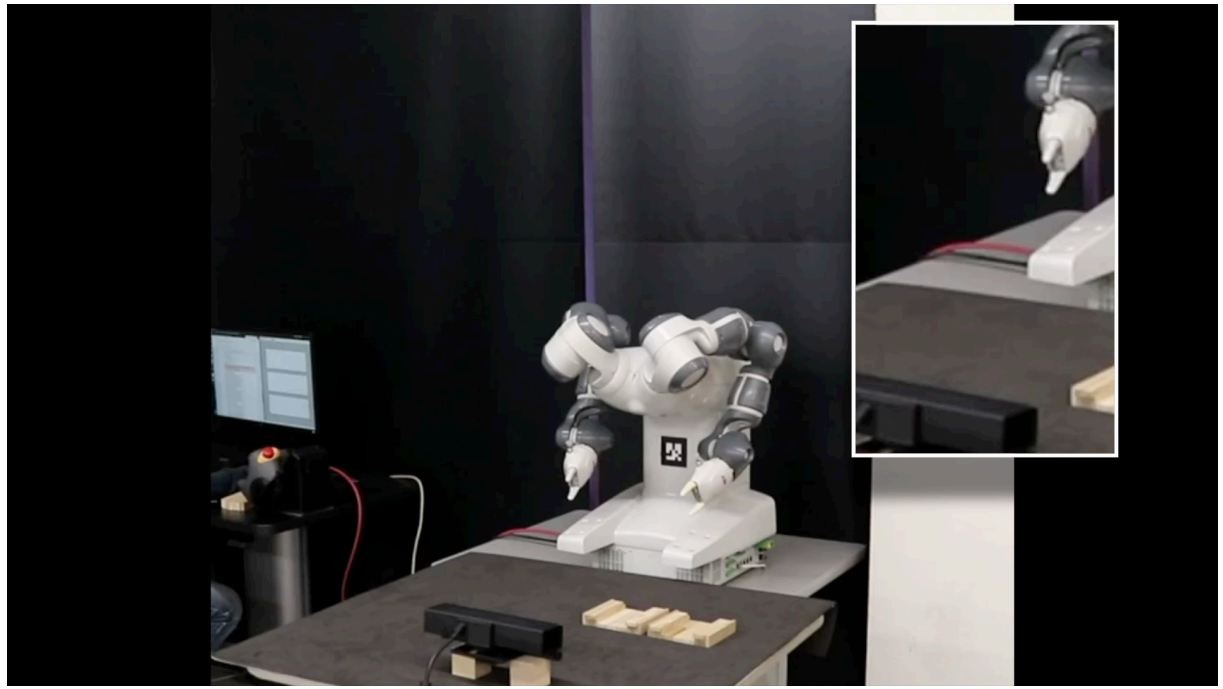
Domain	$\lvert\mathbb{P}\rvert$	$\lvert\mathbb{A}\rvert$	$\lvert\mathcal{Q}_{naive}\rvert$	$\lvert\mathcal{Q}\rvert$	Time/$\mathcal{Q}$ (sec)
gripper	5	3	15×2^5	37	0.14
blocks-world	9	4	36×2^9	92	1.73
elevator	10	4	40×2^{10}	109	5.91
logistics	11	6	66×2^{11}	98	11.62
parking	18	4	72×2^{18}	173	12.01
satellite	17	5	85×2^{17}	127	19.53
openstacks	10	12	120×2^{10}	203	11.28

[AAAI GenPlan 2020]

Power of Abstractions

The same abstractions also help compute anytime task and motion policies for solving complex tasks in stochastic settings





[ICRA 2020 (to appear)]