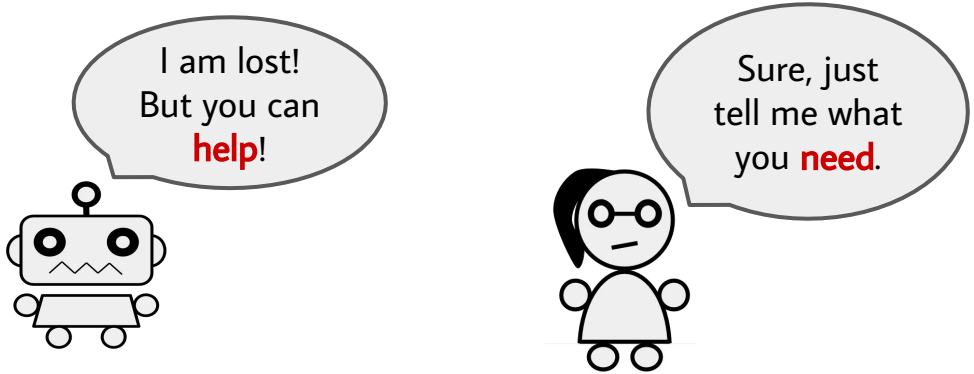
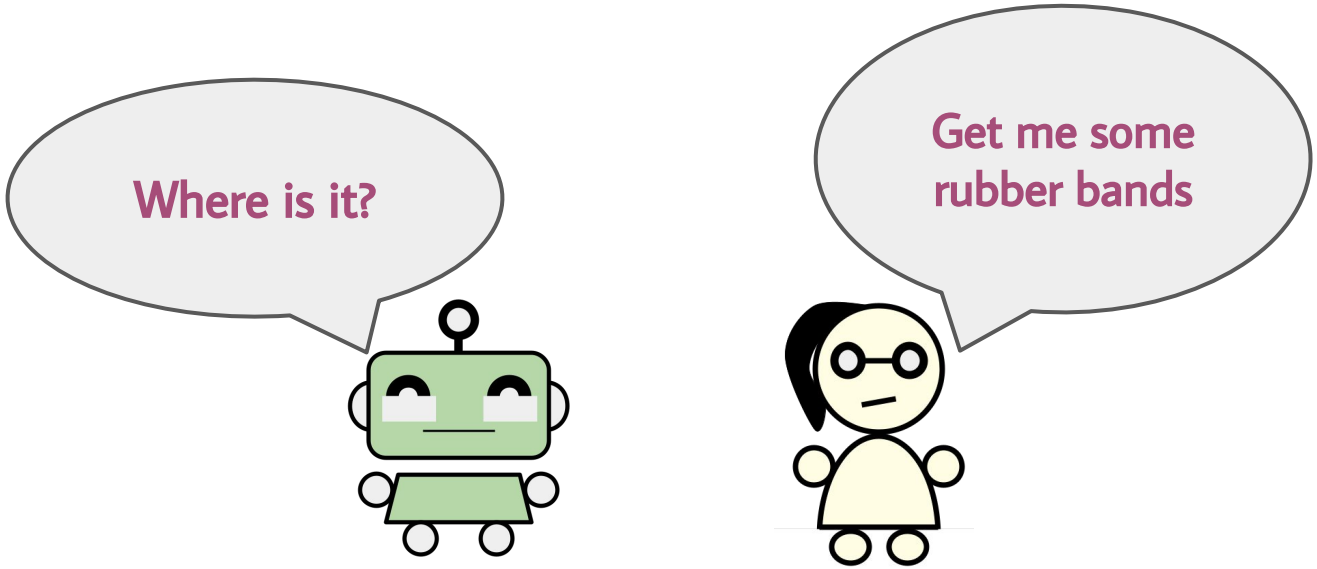
A Framework for Learning to Request Rich and Contextually Useful Information From Humans

Khanh Nguyen *Yonatan Bisk* *Hal Daumé III*



Our contributions



A theoretical framework for *requesting* and *interpreting* information from humans

- How to ask **contextually useful** questions (*speaker* problem)
- How to incorporate **rich** information from human responses (*listener* problem)

Why ask humans? Why bother them?

Machine learning has largely focused on creating agents that can **solve tasks by themselves**

These agents can increase human productivity but offer **limited utility and safety** for regular users

- Limited utility: Capabilities are bounded by training procedure (data, architecture, objective, etc.)
- Limited safety: No natural mechanisms for users to intervene to prevent catastrophic mistakes

Why ask humans? Why bother them?

Machine learning has largely focused on creating agents that can **solve tasks by themselves**

These agents can increase human productivity but offer **limited utility and safety** for regular users

- Limited utility: Capabilities are bounded by training procedure (data, architecture, objective, etc.)
- Limited safety: No natural mechanisms for users to intervene to prevent catastrophic mistakes

Agents that can be assisted by humans are more helpful and safer for humans

- Enhanced utility: human-agent collaboration can be more efficient and effective compared to
  (a) Human does nothing, lets agent screw up the task *(ineffective)*
  (b) Human does the task by themselves *(inefficient)*

- Enhanced safety: agent is aware of its limitations and can convey them to users

Limitations of Current Frameworks

1. Active learning (Angluin 1988; Cohn+, 1994)

   Limited communication protocol

   - Single question type (e.g., asking for a reference label)
   - Primitive information (e.g., reward, low-level action/label)

Limitations of Current Frameworks

1. Active learning (Angluin 1988; Cohn+, 1994)

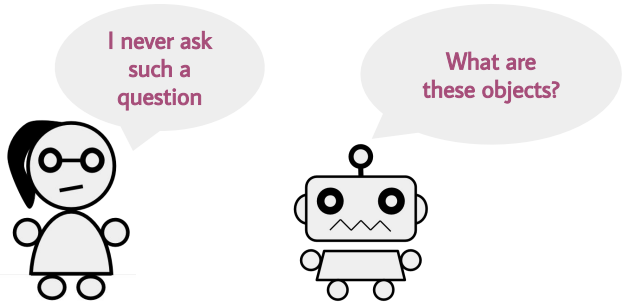
   Limited communication protocol

   - Single question type (e.g., asking for a reference label)
   - Primitive information (e.g., reward, low-level action/label)

2. Imitating human-generated questions (Mostafazadeh+, 2016; De Vries+, 2017; Shi+, 2022)

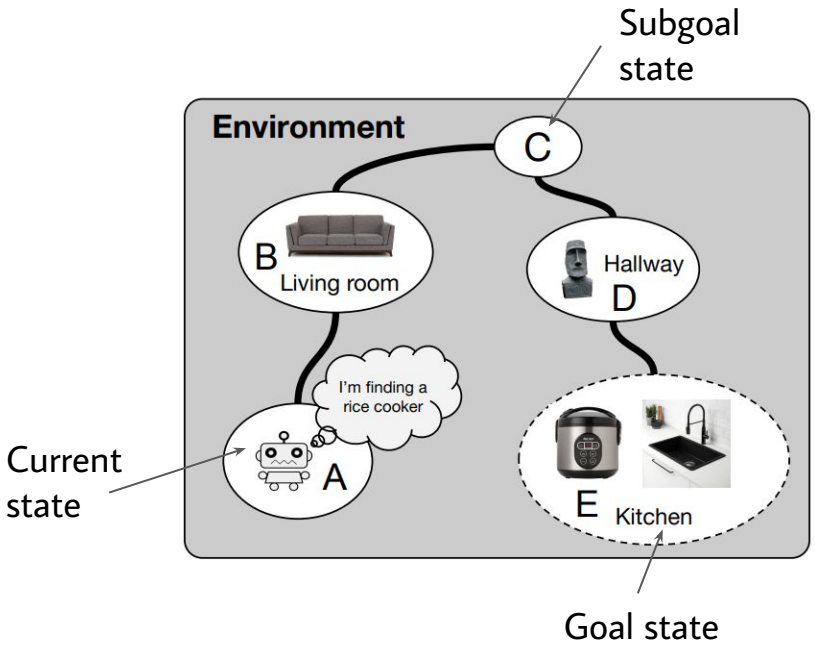
   No emergence of cognitive capabilities

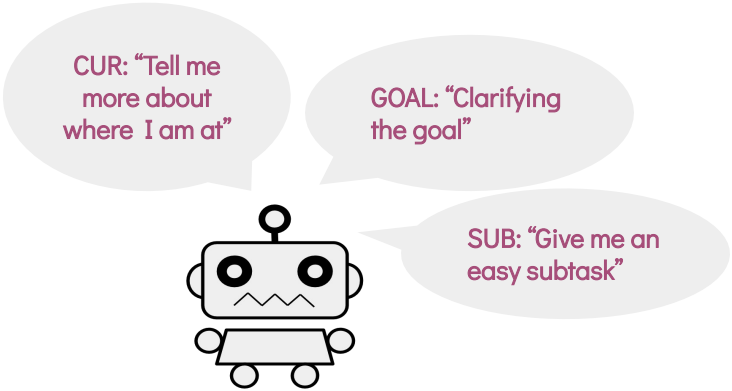
   - Questions asked by a human are helpful for *themselves*, not for agent
   - Agent may not learn to determine what information is useful for *itself*



Speaker Problem: What to Ask?

Equip agent with **specific information-seeking intentions**

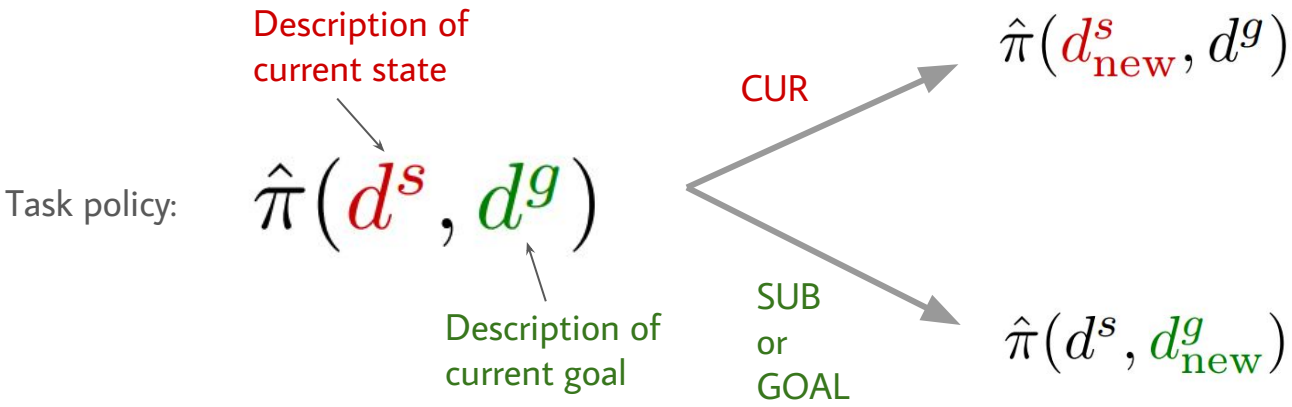




Useful for solving *any* decision-making problem

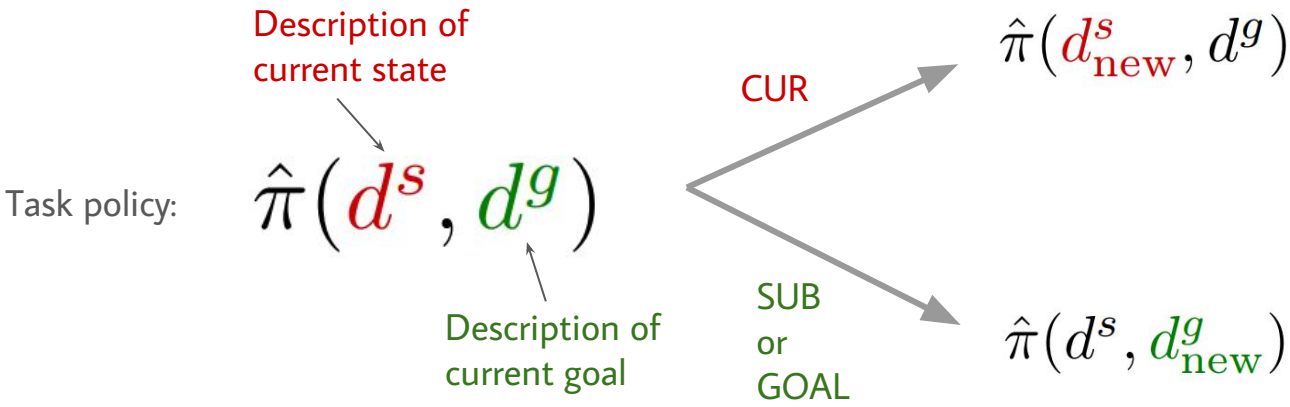
Listener Problem: How to Incorporate Answers?

Upon a request, agent receives a **new description** of the corresponding state



Listener Problem: How to Incorporate Answers?

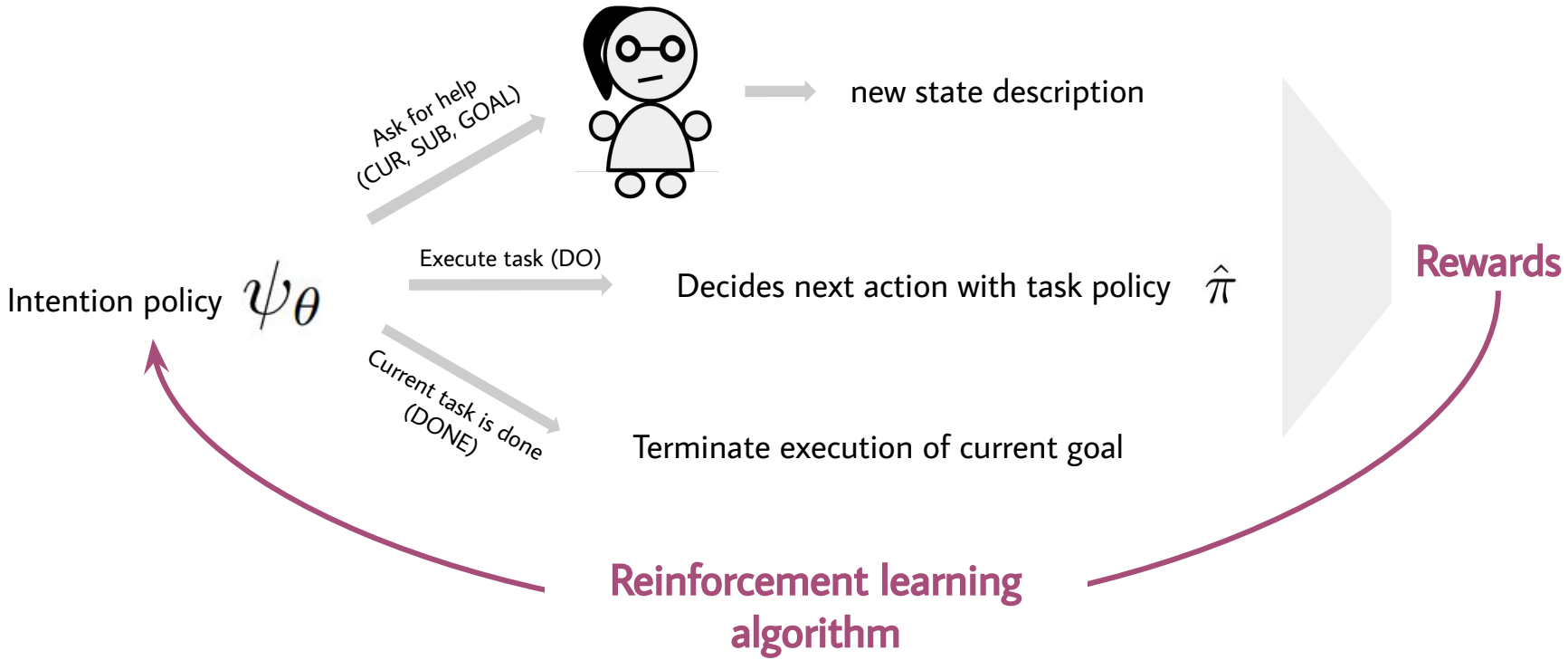
Upon a request, agent receives a **new description** of the corresponding state



Rich, easily extensible communication interface (compared to traditional active learning):

- Humans can provide any information that agent can interpret
- Agent can be trained to interpret new information
- Leverage neural-net architecture to encode various forms of information (e.g. text, images, sequences, etc.)

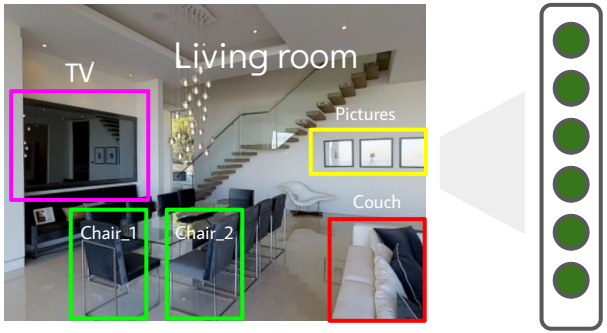
Training-time Behavior

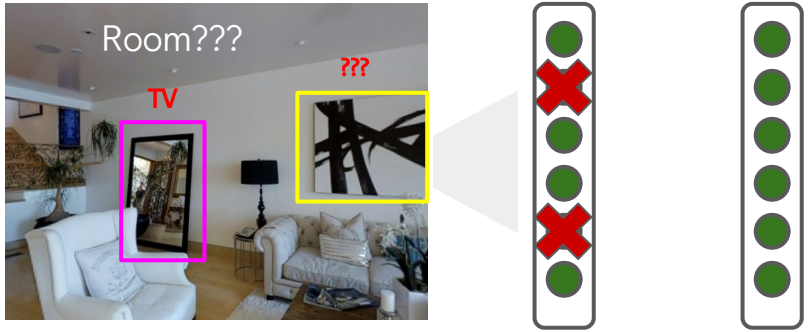


Experiments

Human-assisted navigation tasks: "*Find object O in room R*" (Matterport3D (Anderson et al., 2018))

Emulate **sim-to-real** transfer scenario:





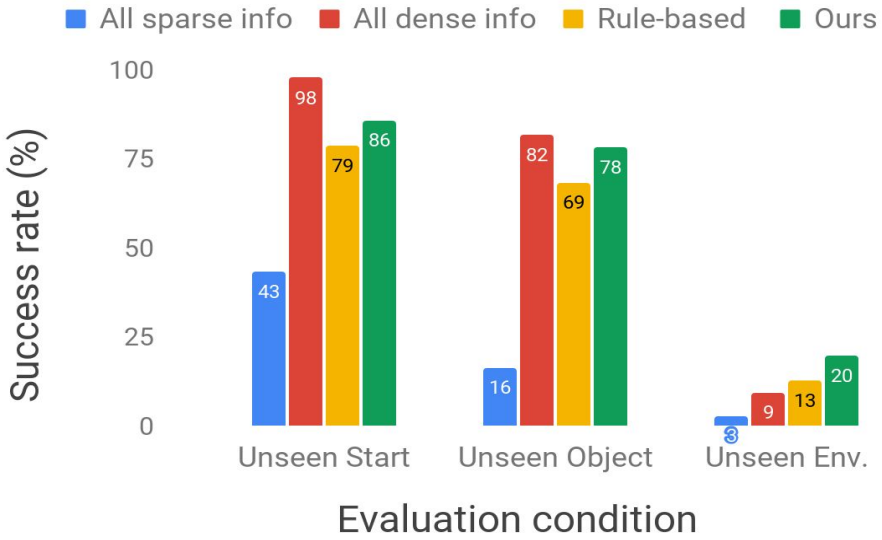
"Sim" condition (perfect information)

- **Dense** feature vectors representing current/goal location
- Features = current room, nearby objects

"Real" condition (perception degrades)

- **Sparse** feature vectors (dropping features from dense representation)
- Can ask a **(simulated) human** for dense feature vectors

Main results

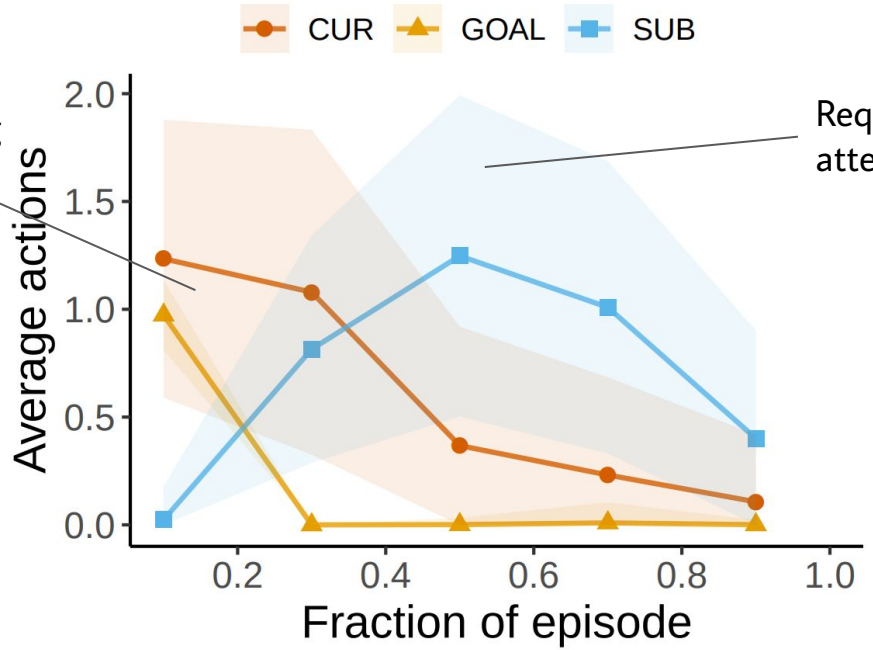


Our agent generalizes better in **novel conditions** thanks to the ability to ask for help

- 2-7x improvement in SR compared to **fully autonomous agent**
- Almost matches or outperforms **agent that always has access to dense information**
- ~¼ of actions are help requests (dense-info agent: always request, autonomous agents: no requests)

Request patterns

Request less information as getting closer to goal



Requesting subgoals **only after** attempting main goal for a while

Summary

The ability to leverage human assistance can make AI agents more helpful and safer for humans

Why more helpful?

- **Proactively** extend its knowledge
- Incorporate **contextually relevant, refined** knowledge given by humans

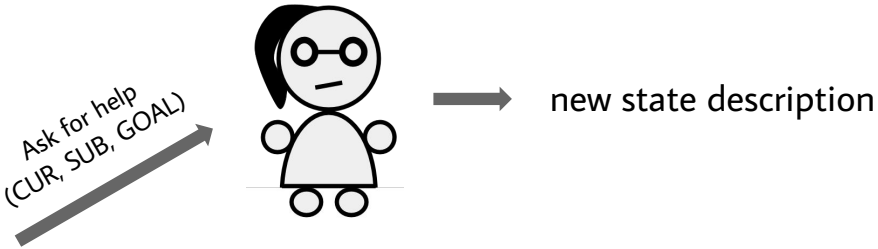
Why safer?

- **Empower** users to influence and control agents **without ML expertise**
- Equip agents with **incentives and capabilities** to convey their limitations and consult humans on difficult decisions

https://github.com/khanhptnk/hari

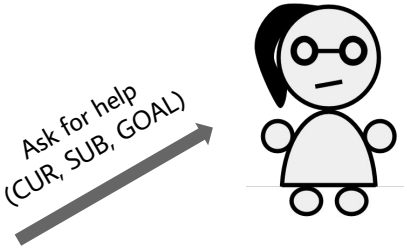
Thank you 😊

Test-time Behavior



Intention policy ψ_θ

Test-time Behavior



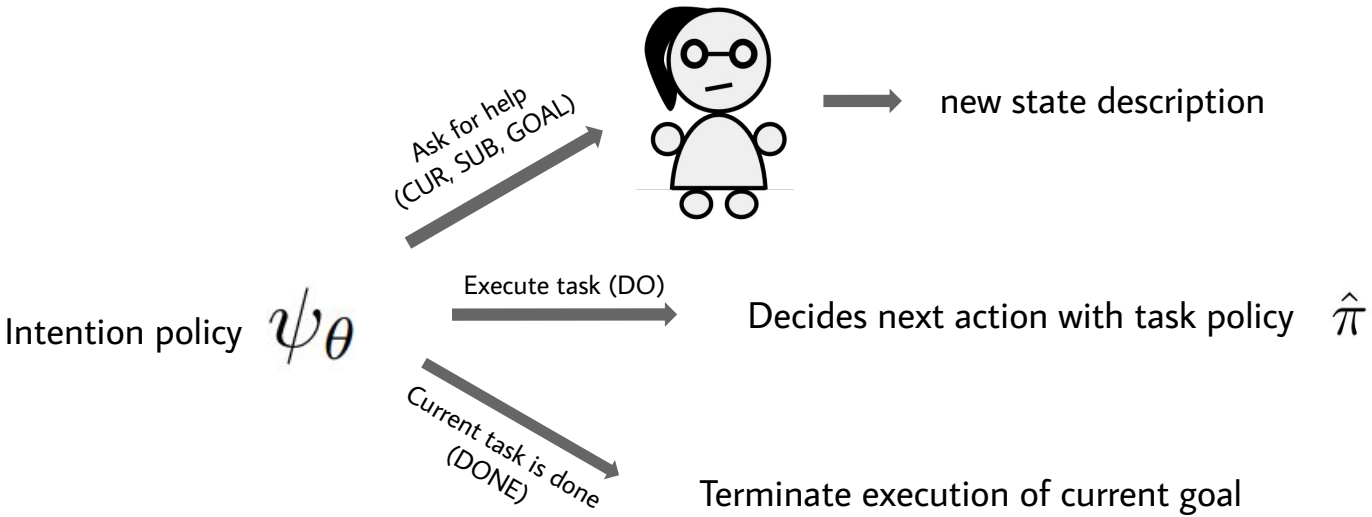
new state description

Intention policy ψ_θ

Execute task (DO)

Decides next action with task policy $\hat{\pi}$

Test-time Behavior



Test-time Behavior