ICAPS 2021 Workshop on Knowledge Engineering for Planning and Scheduling

Learning User-Interpretable Descriptions of Black-Box AI System Capabilities

Pulkit Verma

Shashank Rao Marpally

Siddharth Srivastava

Arizona State University

How Would End Users Assess Their AI Systems?

- How would a lay user determine whether an AI agent will be safe/reliable for a certain task?
- More challenging in settings where agent's internal code is not available (black-box).
- Unproductive usage or safety risks in working with imperfect systems.
- Can we get insights from how we assess humans in such situations?

Do we learn AI system's action descriptions?

Agent Actions (Keystrokes)	Learned Capabilities
W	kill_monster
A	goto_door
S	goto_key
D	goto_monster
E	pick_key
	open_door

Knowledge of primitive actions might be insufficient to understand the agent's capabilities

User-vocabulary may be limited

Agent's State Representation	Interpretable State Representation
pixel_1_1(#42A8B3)	monster_at(5,3)
pixel_1_2(#42A8B3)	player_at(6,3)
.	key_at(9,4)
.	door_at(9,2)
.	
pixel_n_m(#203A3D)	

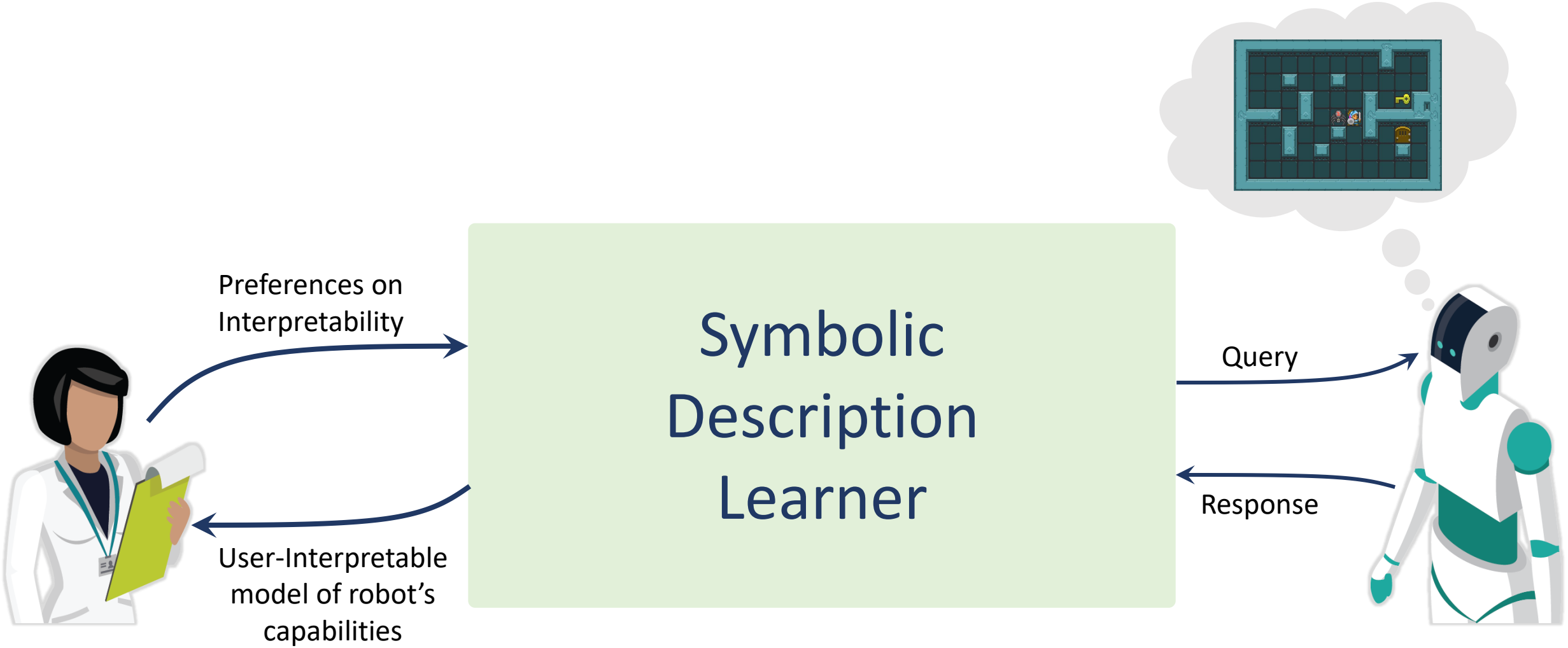
Might be more expressive than what the user understands

Related Work

Learning high-level symbolic models of AI systems using observations or interventions.

- Not interpretable, assume access to predefined options: Konidaris et al. (JAIR'18)
- Assume precise user-vocabulary: AIA - Verma et al. (AAAI'21)
- Needs hand-coding of states: Zhang et al. (ICML'18)
- Require lot of data: Schema Networks - Kansky et al. (ICML'17), Agarwal et al. (NIPS'16)
- Use passive observations: LOCM - Cresswell et al. (ICAPS'09), ARMS -Yang et al. (AIJ 2007), LOUGA - Kučera and Barták (KMAIS 2018), FAMA - Aineto et al. (AIJ 2019)

Our Approach: Symbolic Description Learner

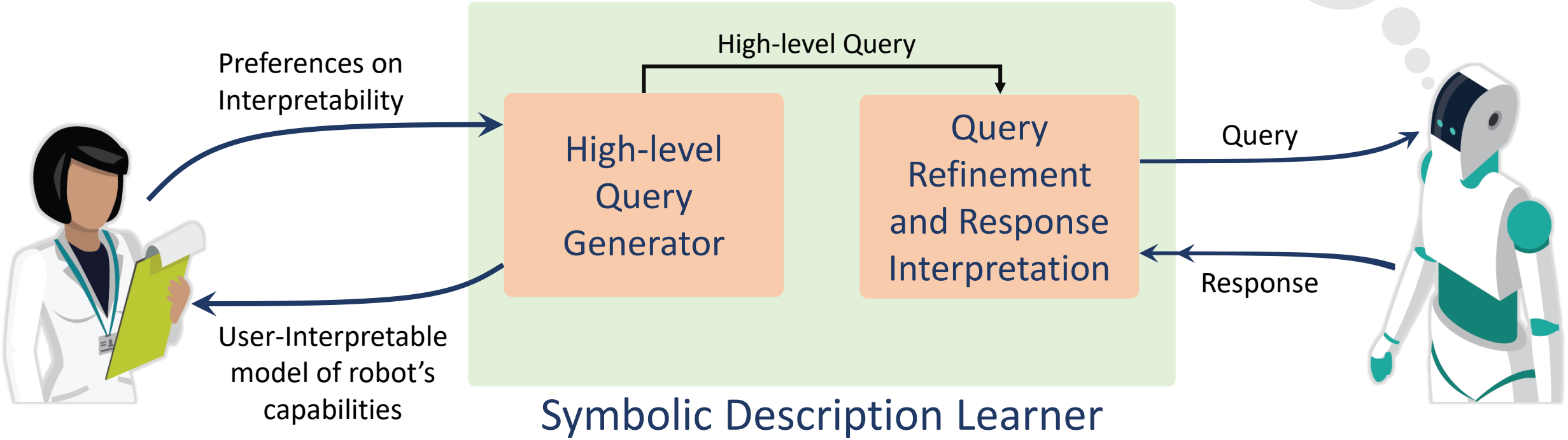


Our Approach: Symbolic Description Learner

In terms that the user understands.
Query: Initial State and Plan
Response: Length of plan that can be executed successfully and the final state

What do you think will happen if you don't have key and you move to a location near the door and then open the door?

I can execute only the first step. After this I will still be in a location near the door.

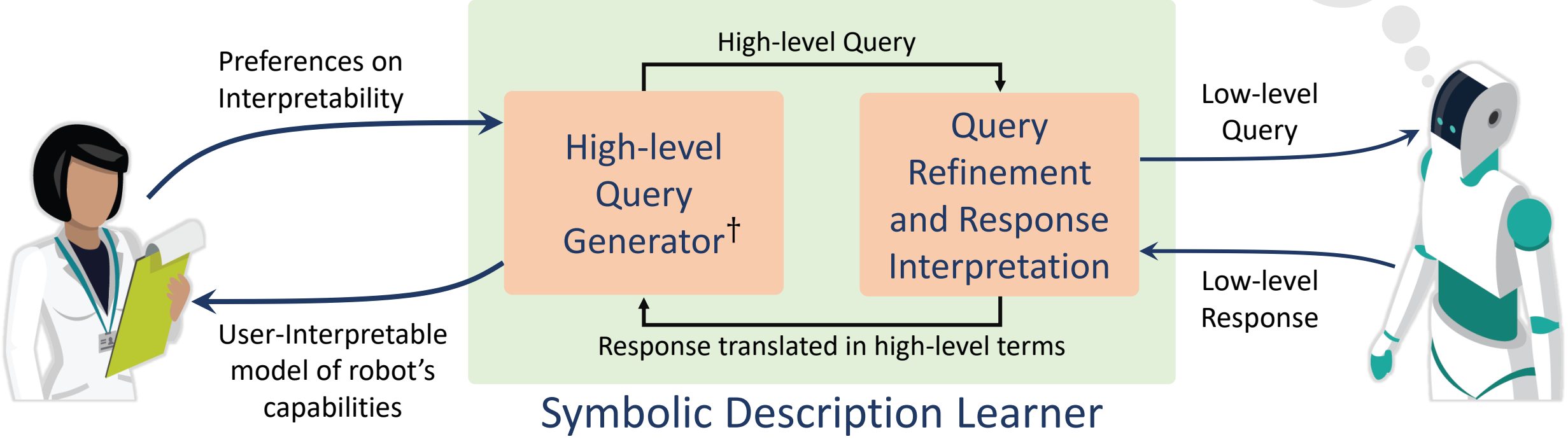


Our Approach: Symbolic Description Learner

In terms that the agent uses.
Query: Initial State and Goal State
Response: Yes/No. Representing if it can reach from the initial state to the goal state.

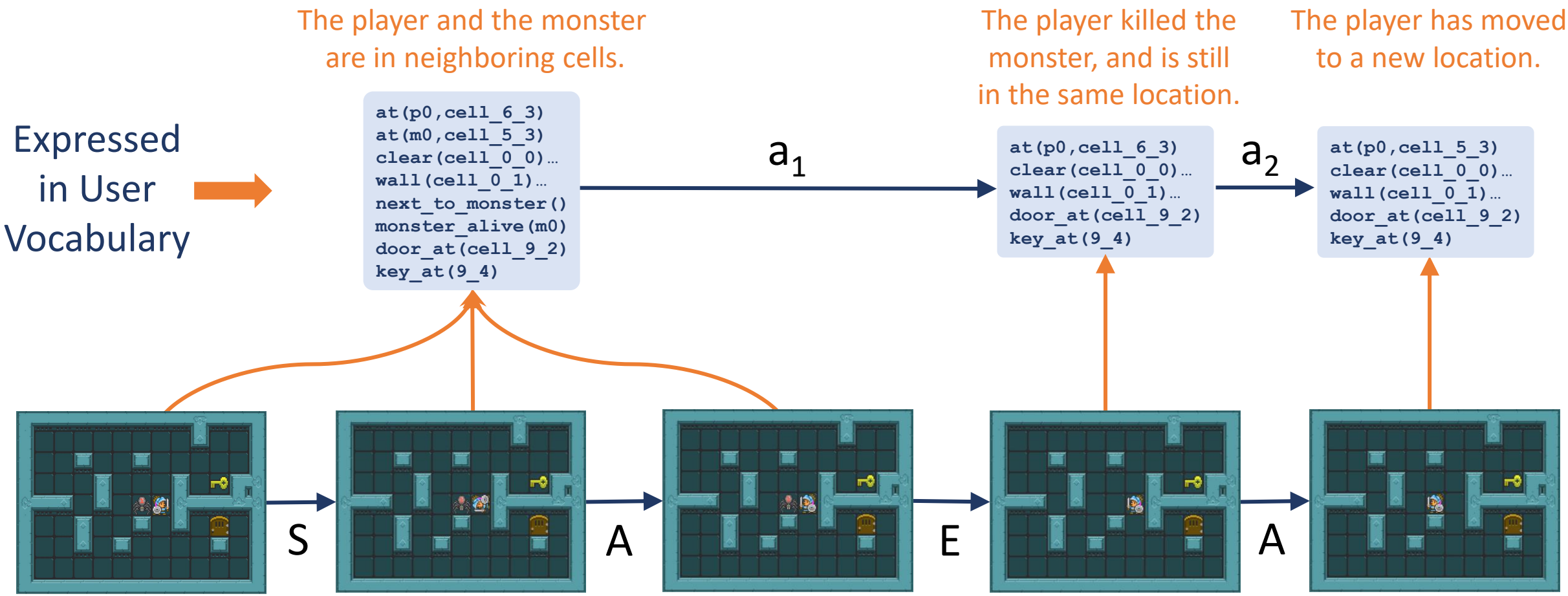
Can you reach from state <initial_state> to <goal_state>?

Yes/No



†Verma, P.; Marpally S. R.; and Srivastava, S. Asking the Right Questions: Learning Interpretable Action Models through Query Answering. In AAAI 2021.

Temporal Abstraction



Experimental Setup

- Randomly generate an environment from one of four GVGAI Games.
- Initialize two kinds of agents –
  - Search Agent: Use search algorithms.
  - Policy Agent: Use black-box policies.
- Vary grid size to see variations in number of queries and time taken per query.

Results

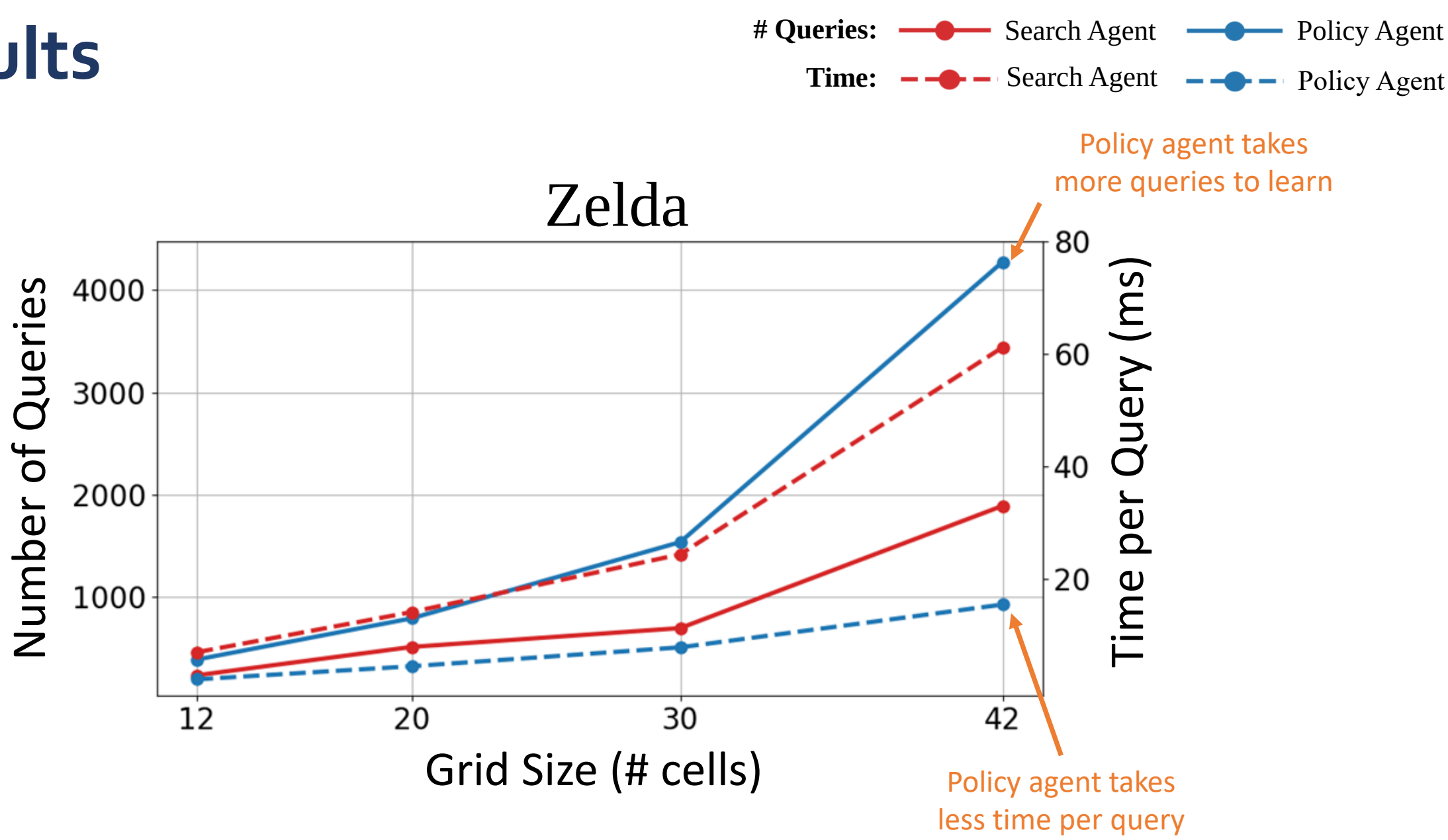
Some of the actions learned for Zelda

```
(:action kill-monster
 :parameters ()
 :precondition (and
    (at p1 6-3)
    (at m1 5-3)
    (monster_alive m1)
    (next_to_monster))
 :effect (and
    (clear 5-3)
    (not (at m1 5-3))
    (not (monster_alive m1))
    (not (next_to_monster))))
```

```
(:action get-key
 :parameters ()
 :precondition (and
    (at p1 1-2)
    (at key 0-2))
 :effect (and
    (not (at p1 1-2))
    (not (at key 0-2)
    (clear 1-2)
    (has-key)
    (at p1 0-2)))
```

```
(:action escape-door
 :parameters ()
 :precondition (and
    (at p1 1-1)
    (at door 2-1)
    (clear 2-1)
    (has_key)
    (not (monster_alive m1)))
 :effect (and
    (not (at p1 1-1))
    (not (clear 2-1))
    (clear 1-1)
    (escaped)
    (at p1 2-1)))
```

Results


Queries: Search Agent Policy Agent
Time: Search Agent Policy Agent
Policy agent takes more queries to learn
Zelda
Number of Queries
4000
3000
2000
1000
80
60
40
20
Time per Query (ms)
12
20
30
42
Grid Size (# cells)
Policy agent takes less time per query

Conclusions

The proposed approach:

- Efficiently learns internal model of an agent in a STRIPS-like form.
- Needs no prior knowledge of the agent model.
- Only requires an agent to have rudimentary query answering capabilities.
- Learns the model in terms of concepts that the user understands.

Paper

bit.ly/36VABk3

verma.pulkit@asu.edu