CHAI - 6th Annual Workshop, 2022

Discovering User-Interpretable Capabilities of Black-Box AI Agents*

Pulkit Verma

Arizona State University

Center for Human-Compatible Artificial Intelligence

AAIR Autonomous Agents and Intelligent Robots

ASU Arizona State University

*Based on “Verma, P.; Marpally, S. R.; Srivastava S. *Discovering User-Interpretable Capabilities of Black-Box Planning Agents*. In Proc. KR 2022.”

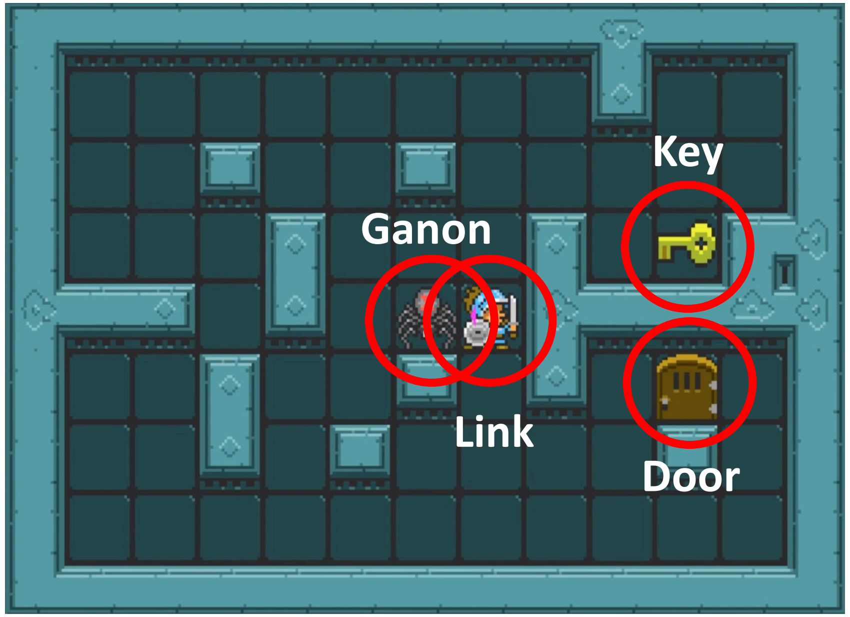
Personalized Assessment of Taskable AI Systems

- AI systems should make it easy for its operators to learn how to use them safely.[†]
- Users can give them multiple tasks.
  - How would users know what they can do?
- Should work with black-box AI systems.

[†] Srivastava S. *Unifying Principles and Metrics for Safe and Assistive AI.* In Proc. AAAI 2021.

Capability v/s Functionality

- *Functionality*: Set of possible low-level actions of the agent.
- *Capability*: What agent's planning and learning algorithms can do.



Agent Actions
(Keystrokes)

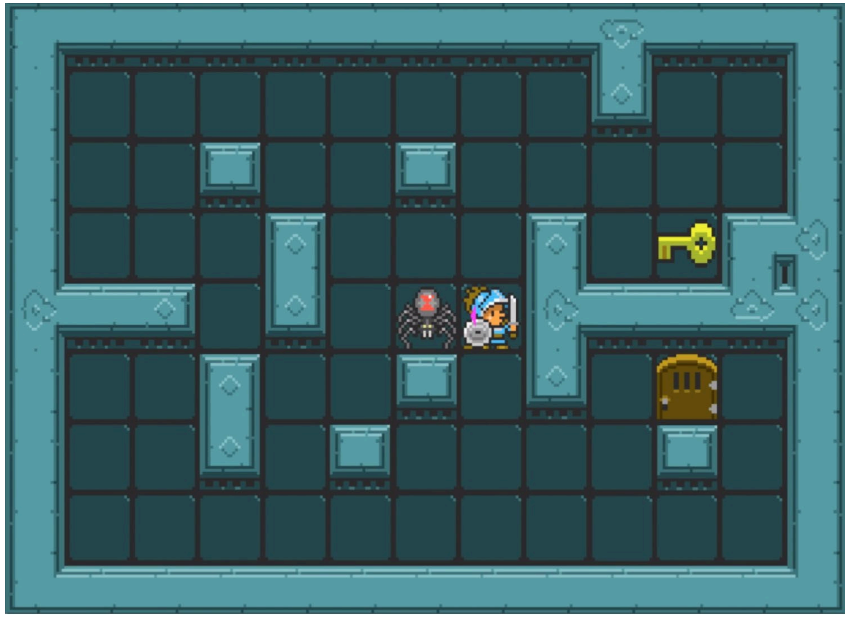
W
A
S
D
E

Learned
Capabilities

(defeat ganon)
(go to door)
(go to key)
(go to ganon)
(pick key)
(open door)

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

User-vocabulary may be limited

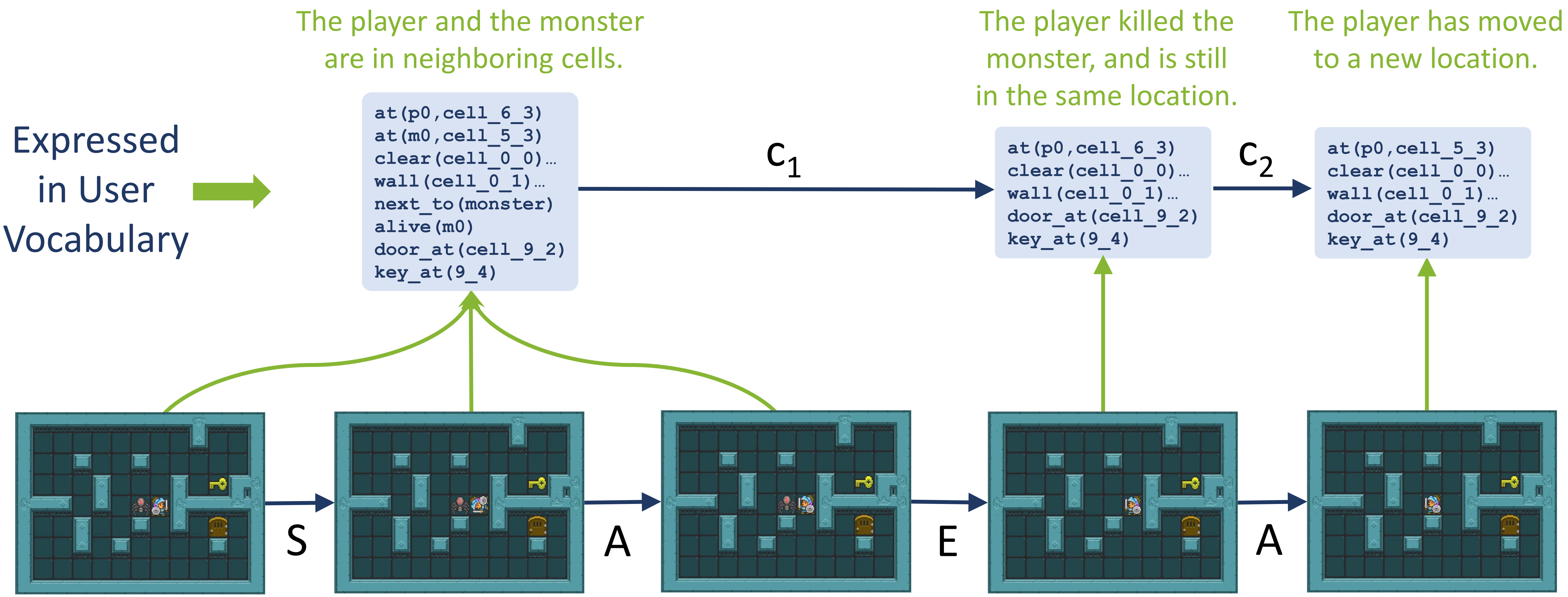


Agent’s State Representation	Interpretable State Representation

pixel_1_1(#42A8B3)
pixel_1_2(#42A8B3)
.
.
.
pixel_n_m(#203A3D)
	(at ganon 5,3)
	(at link 6,3)
	(at key 9,4)
	(at door 9,2)

Might be more expressive than what the user understands

Discovering Capabilities



Expressed in User Vocabulary

The player and the monster are in neighboring cells.

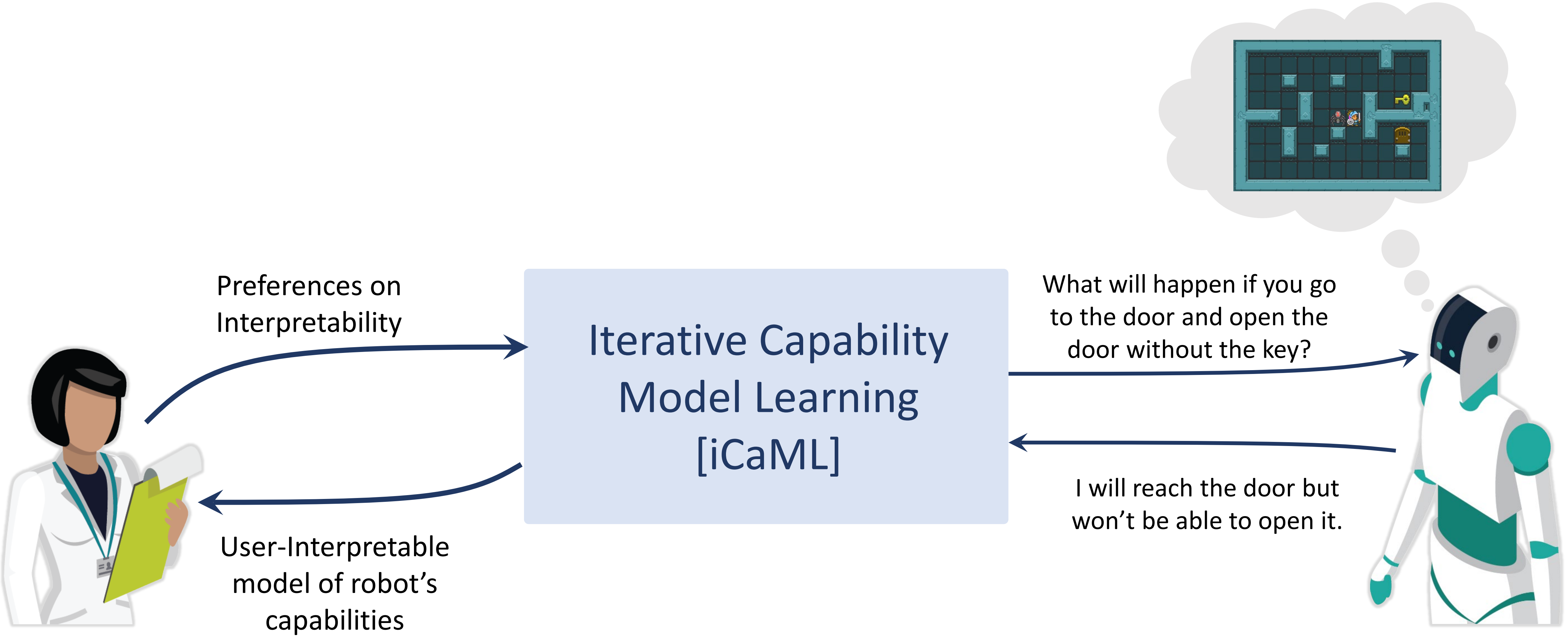
```
at(p0,cell_6_3)
at(m0,cell_5_3)
clear(cell_0_0)…
wall(cell_0_1)…
next_to(monster)
alive(m0)
door_at(cell_9_2)
key_at(9_4)
```

The player killed the monster, and is still in the same location.

```
at(p0,cell_6_3)
clear(cell_0_0)…
wall(cell_0_1)…
door_at(cell_9_2)
key_at(9_4)
```

The player has moved to a new location.

```
at(p0,cell_5_3)
clear(cell_0_0)…
wall(cell_0_1)…
door_at(cell_9_2)
key_at(9_4)
```



Preferences on Interpretability

Iterative Capability Model Learning [iCaML]

What will happen if you go to the door and open the door without the key?

I will reach the door but won’t be able to open it.

User-Interpretable model of robot’s capabilities

Results: Example of Learned Capability

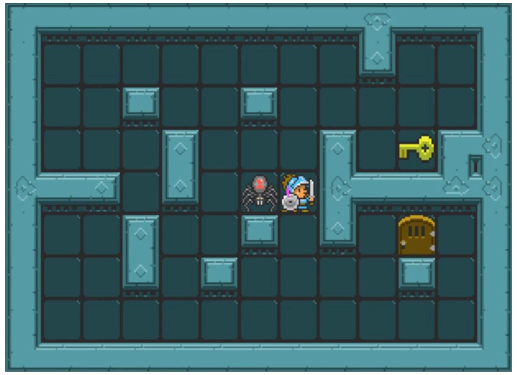
```
(:capability c4
 :parameters (?player1 ?cell1
   ?monster1 ?cell2)
 :precondition
  (and (alive ?monster1)
    (at ?player1 ?cell1)
    (at ?monster1 ?cell2)
    (next_to ?monster1))
 :effect
  (and (clear ?cell2)
   (not(alive ?monster1))
   (not(at ?monster1 ?cell2))
   (not(next_to ?monster1))))
```

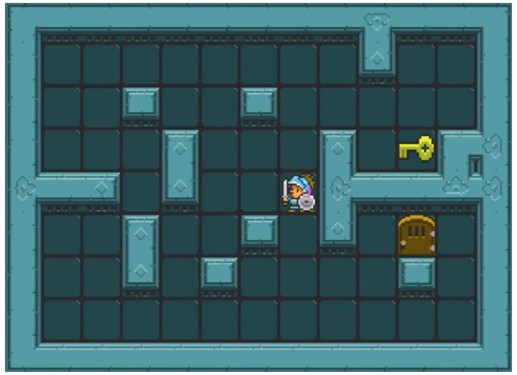
Position of Link has not changed

Ganon is not at its previous location

Ganon is not alive anymore

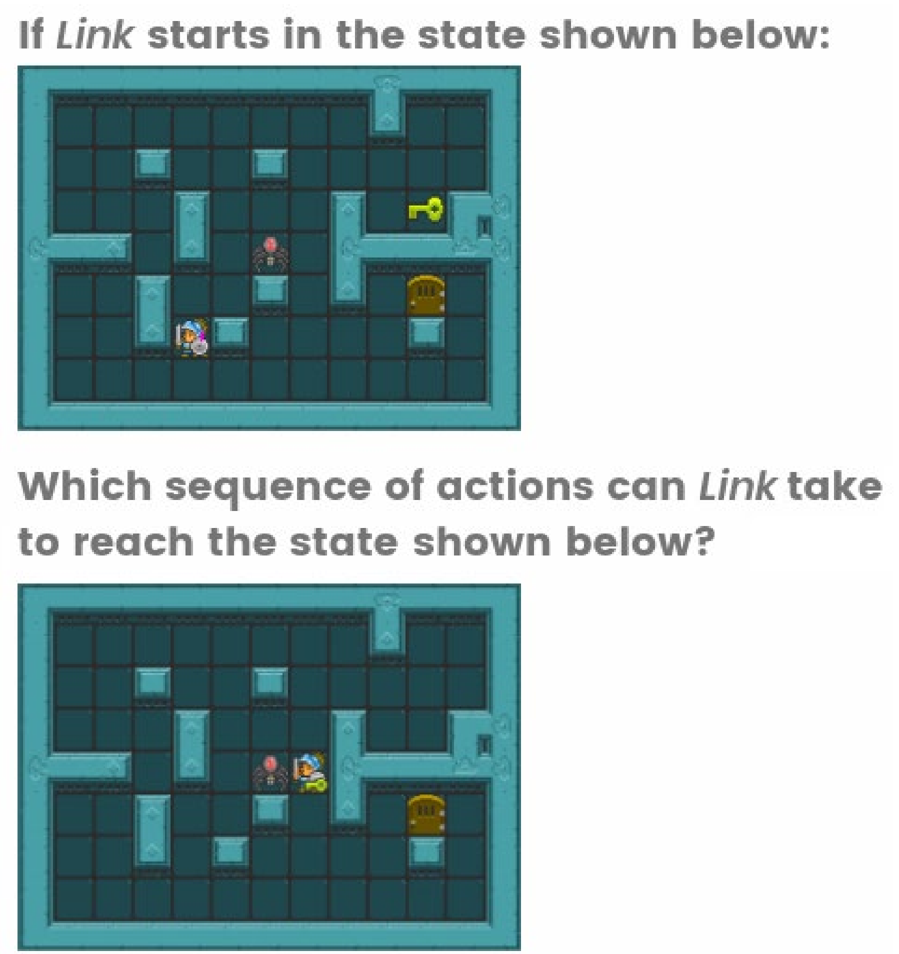
Link is not next to Ganon





Behavior Evaluation Study

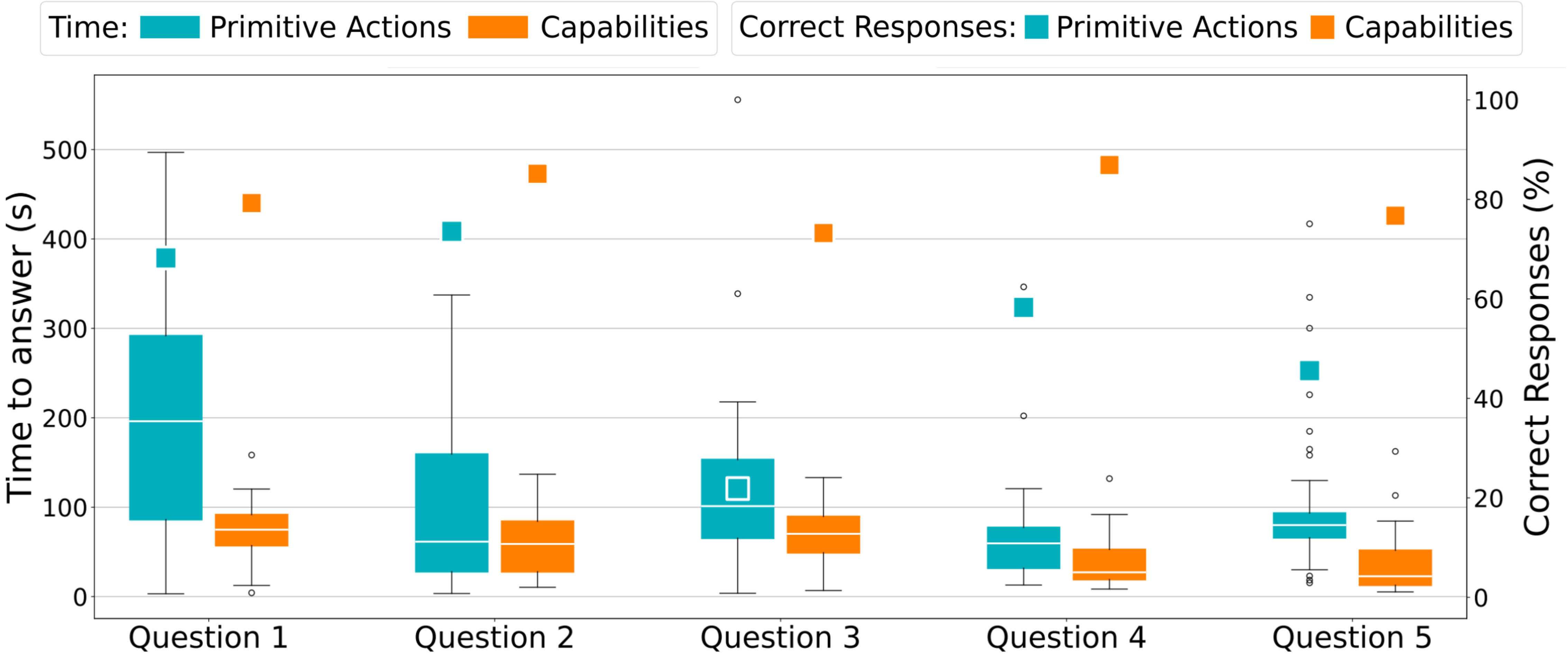
- Rules of Zelda-like game explained to users.
- 108 participants split into 2 groups of 54 each.
- One group shown action descriptions in English, another group shown capability descriptions.
- Asked to answer 5 questions like shown here.



If *Link* starts in the state shown below:

Which sequence of actions can *Link* take to reach the state shown below?

Results: Behavior Evaluation Study



Formal Results

- The learned descriptions are consistent with the observations.
- This approach is maximally consistent, i.e., we cannot add any more literals to the preconditions or effects without ruling out some truly possible models.
- Learned capabilities are realizable, i.e., downward refinement is ensured.
- If a high-level model is expressible deterministically using the user vocabulary and local connectivity holds, then in the limit of infinite execution traces, the probability of discovering all capabilities expressible in the user vocabulary is 1.

Key Takeaways

The proposed approach:

- Efficiently discovers capabilities of an agent in a STRIPS-like form in fully observable and deterministic settings.
- Needs no prior knowledge of the agent model.
- Only requires an agent to have rudimentary query answering capabilities.
- Learns a maximally consistent capability model accurately with a small number of queries.

Paper

arxiv: 2107.13668

🌐 pulkitverma.net

✉ verma.pulkit@asu.edu