



IIT Ropar | Research Talk

October 19, 2024

Learning Interpretable Models for Personalized Assessment of Taskable AI Systems

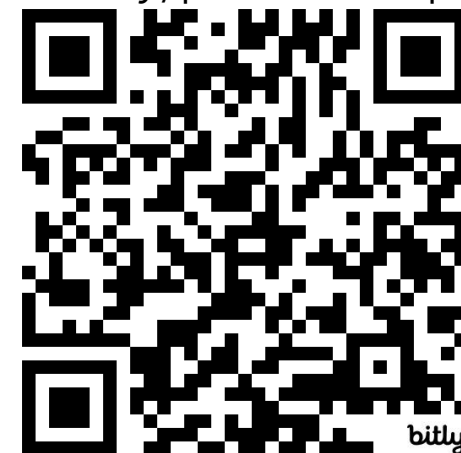
Pulkit Verma

✉ pulkitv@mit.edu

🌐 pulkitverma.net



bit.ly/pulkit-iitrpr



01

Research Area &
Interesting Problems

02

Agent Interrogation
Algorithm
(AAAI 2021)

03

Future Research
Directions

04

Publications

05

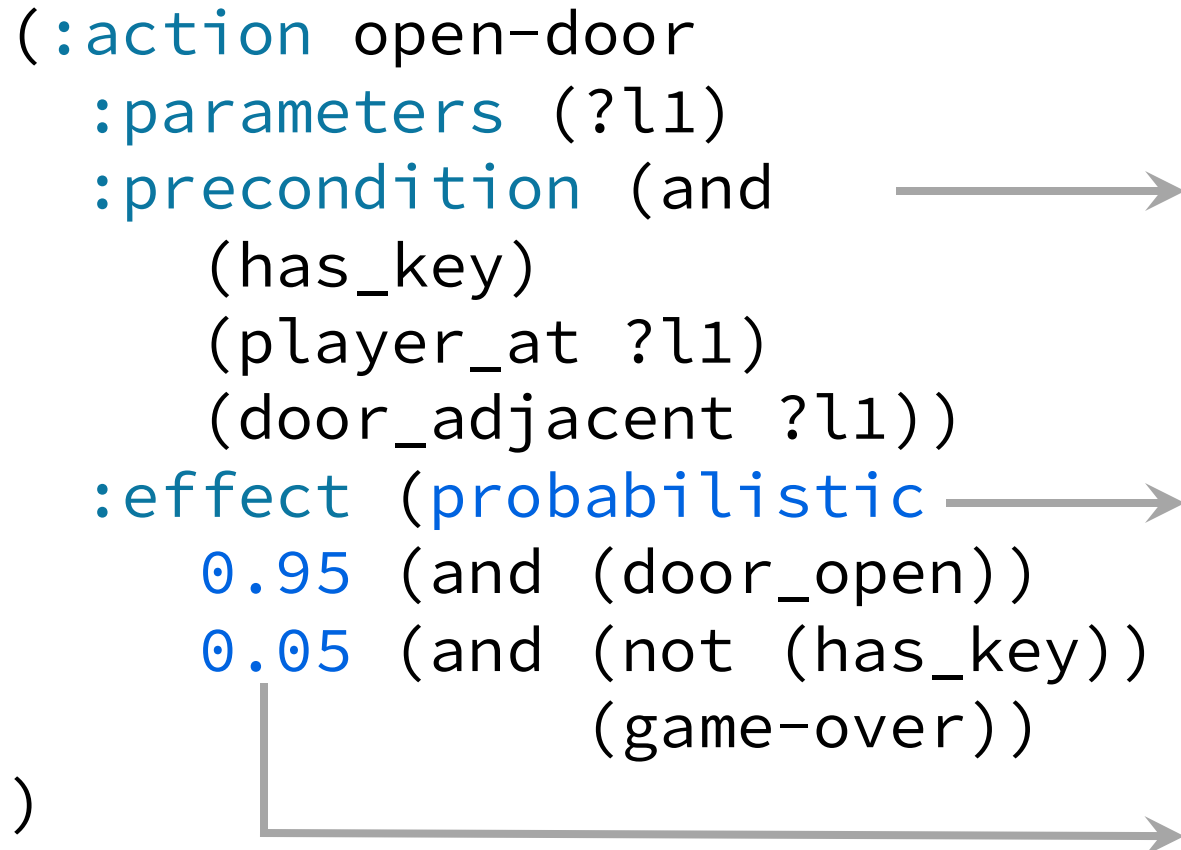
Long Term
Research Plan

Automated Planning

- To complete any task, we plan using a sequence of actions.
- *Automated Planning:*
 - Has discrete high-level actions.
 - Each high-level action corresponds to a sequence of low-level actions or controls.
 - Given an initial state and a goal, what sequence of actions, called a plan, should an AI system execute to achieve the goal?

Probabilistic Planning Domain Definition Language

```
(:action open-door
  :parameters (?l1)
  :precondition (and
    (has_key)
    (player_at ?l1)
    (door_adjacent ?l1))
  :effect (probabilistic
    0.95 (and (door_open))
    0.05 (and (not (has_key))
              (game-over))
  )
```



Precondition: This condition must be true for this action to execute

Effect: This is a set of conditions, one of which becomes true when this action is executed

Probabilities: Each set of effect has an associated probability with which that effect set is executed

Interpretable: Easily Convertible to Natural Language

```
(:action open-door
  :parameters (?l1)
  :precondition (and
    (has_key)
    (player_at ?l1)
    (door_adjacent ?l1))
  :effect (probabilistic
    0.95 (and (door_open))
    0.05 (and (not(has_key))
              (game-over))
  )
```

The player can open the door when in location ?l1 if:

- It has the key
- The player is at location ?l1
- The door is adjacent to location ?l1

After executing that capability:

- With 95% probability, the door will open
- With 5% probability, the player will not have the key and the game will be over

Personalized Assessment of AI Systems that can Learn and Plan

- Users would like to give AI systems multiple tasks.
 - How would users know what the AI systems can do?
- AI systems should support third-party assessment.
- The assessment should work with:
 - **Adaptive** AI systems.
 - **Black-Box** AI systems.



```
at(p0,cell_6_3)
clear(cell_0_0)
door_at(cell_9_2)
next_to(m0)
alive(m0)
key_at(9_4)
```

[Input]
Concepts that the
user understands



Personalized
AI-Assessment
Module (AAM)



Arbitrary internal
implementation

Doesn't know
user's vocabulary

Black-Box AI

```
at(p0,cell_6_3)
clear(cell_0_0)
door_at(cell_9_2)
next_to(m0)
alive(m0)
key_at(9_4)
```

[Input]
Concepts that the
user understands



Personalized
AI-Assessment
Module (AAM)

Query

Response



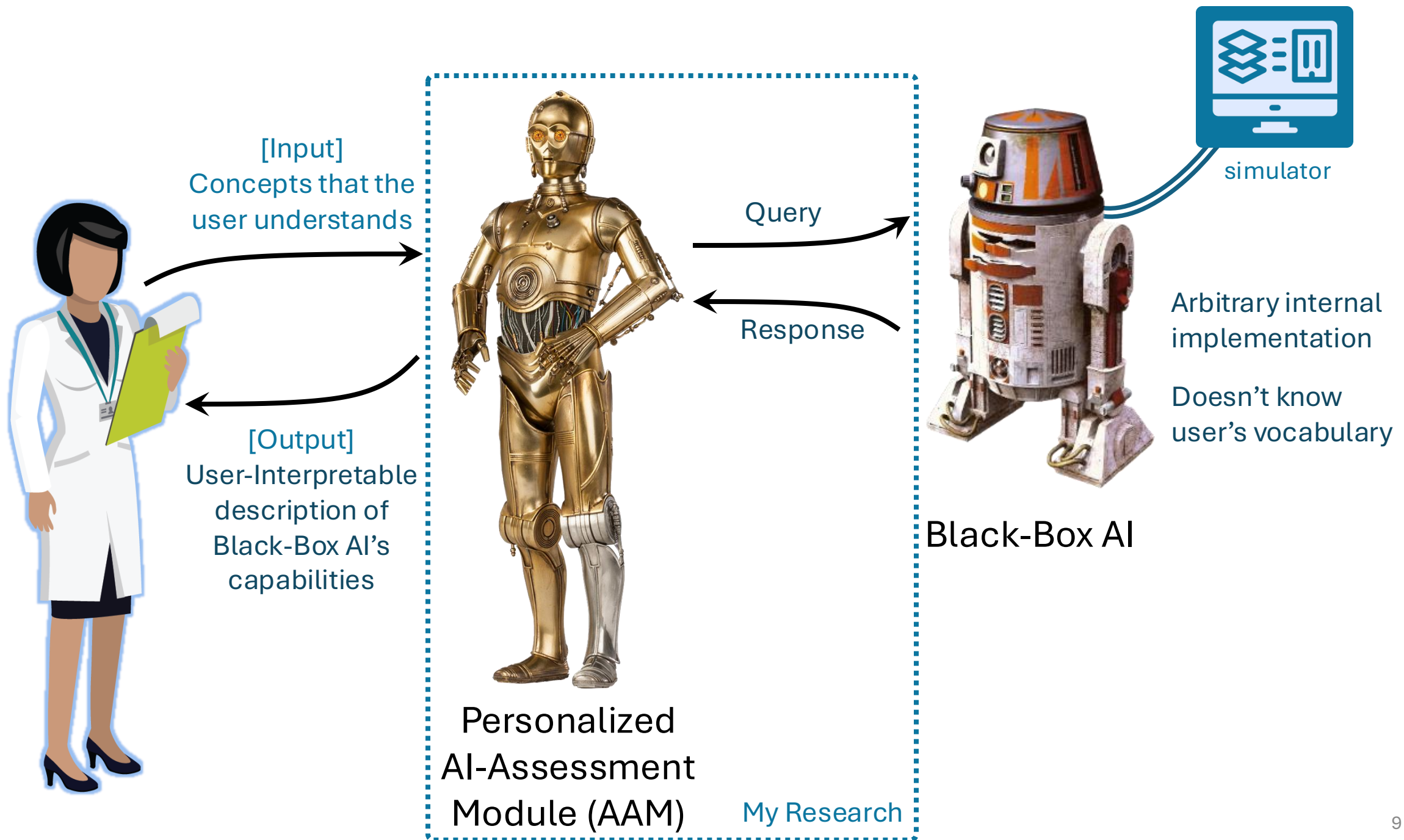
Black-Box AI



simulator

Arbitrary internal
implementation

Doesn't know
user's vocabulary



01

Research Area &
Interesting Problems

02

Agent Interrogation
Algorithm
(AAAI 2021)

03

Future Research
Directions

04

Publications

05

Long Term
Research Plan

Deterministic and Stationary Setting

Input

- Predicates (User vocabulary)
 - With their evaluation functions
- List of capabilities.



Output

- PDDL-like description of each capability.

Assumptions

- User's vocabulary matches simulator's vocabulary.
- Black-Box AI provides a list of capabilities.
- Stationary agent model.
- Deterministic environment.
- Fully observable setting.



Siddharth
Srivastava



Shashank
R Marpally

Asking the Right Questions: Learning Interpretable
Action Models Through Query Answering

Pulkit Verma, Shashank Rao Marpally, Siddharth Srivastava

In Proceedings of the Thirty-Fifth AAAI Conference on
Artificial Intelligence, 2021 (CORE Ranking: A*)

Exponential Search for Learning Correct Description

- Consider the following 4 predicates/concepts:
 - (has_key)
 - (door_open)
 - (door_adjacent ?x)
 - (player_at ?x)
- Consider just one capability:
(open-door ?x)
- $9^{|C| \times |P|} = 9^{1 \times 4} = 6561$ possible models
(Assuming deterministic models/
descriptions, i.e., no probabilities).

```
(:action open-door
  :parameters (?l1)
  :precondition (and
    (+/-/∅) (has_key)
    (+/-/∅) (door_open)
    (+/-/∅) (door_adjacent ?l1)
    (+/-/∅) (player_at ?l1))
  :effect (and
    (+/-/∅) (has_key)
    (+/-/∅) (door_open)
    (+/-/∅) (door_adjacent ?l1)
    (+/-/∅) (player_at ?l1)))
```

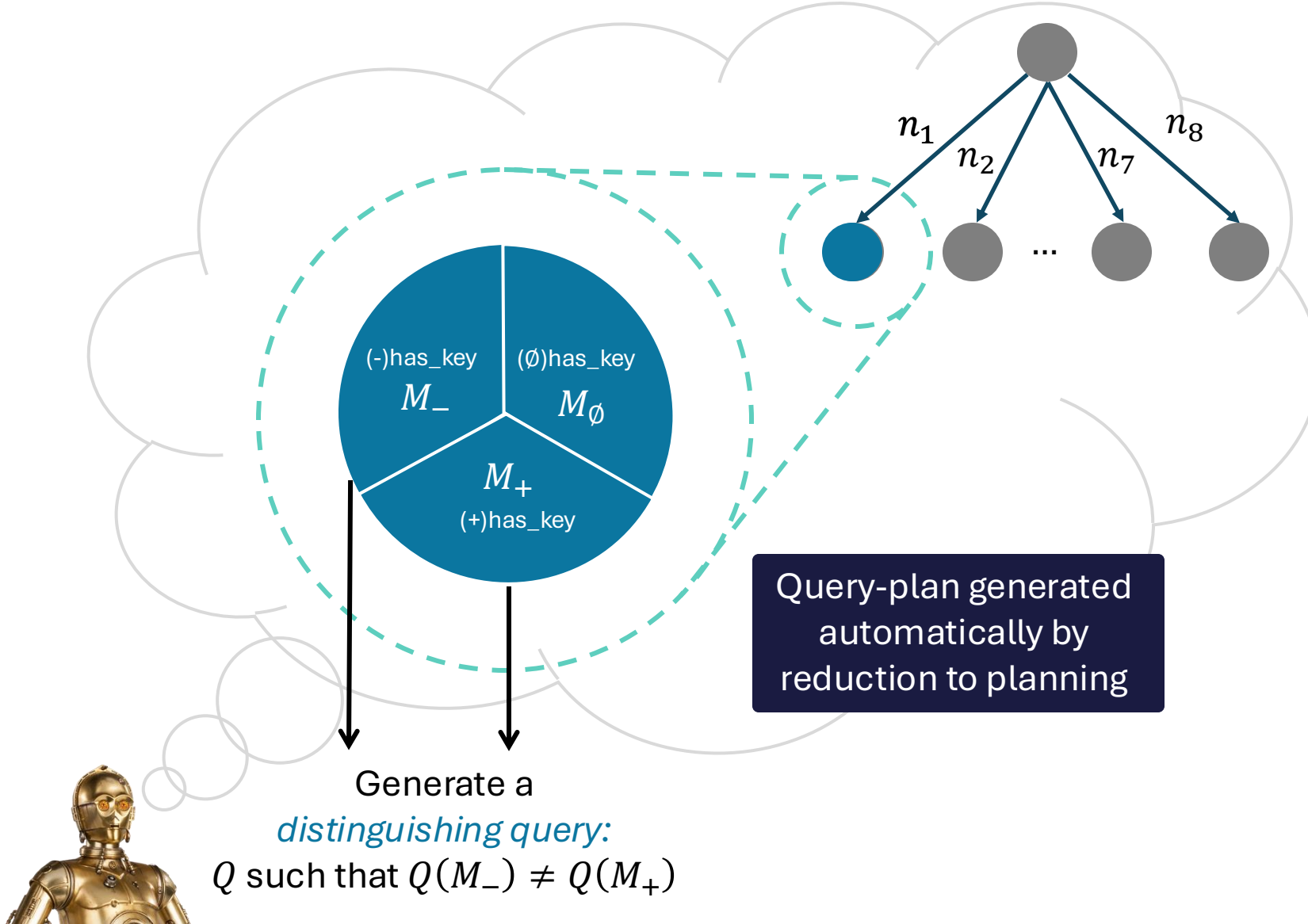
Simple Queries

Query	In state s_I , what will happen if you execute the plan $\pi = \langle c_1, \dots, c_n \rangle$?	Can you go from state s_I to state s_F ?
Response	I can execute first ℓ steps of the plan, ending up in state s_F .	Yes / No.
	Plan Outcome Queries	State Reachability Query

- How to generate the queries?
- How to use the responses to generate models?

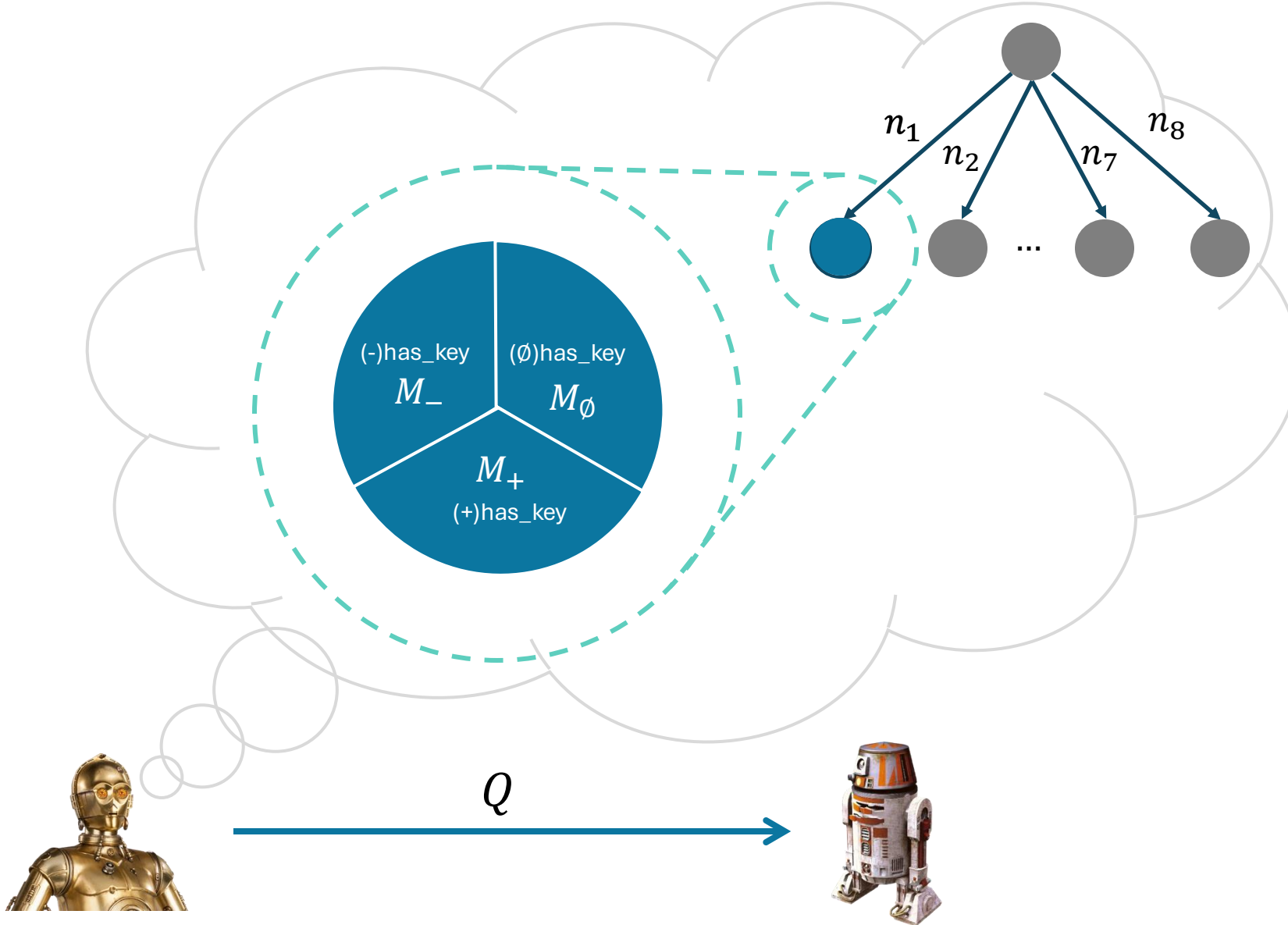
We have a reduction that converts this to a planning problem, so it automatically generates queries.

Hierarchical Query Synthesis



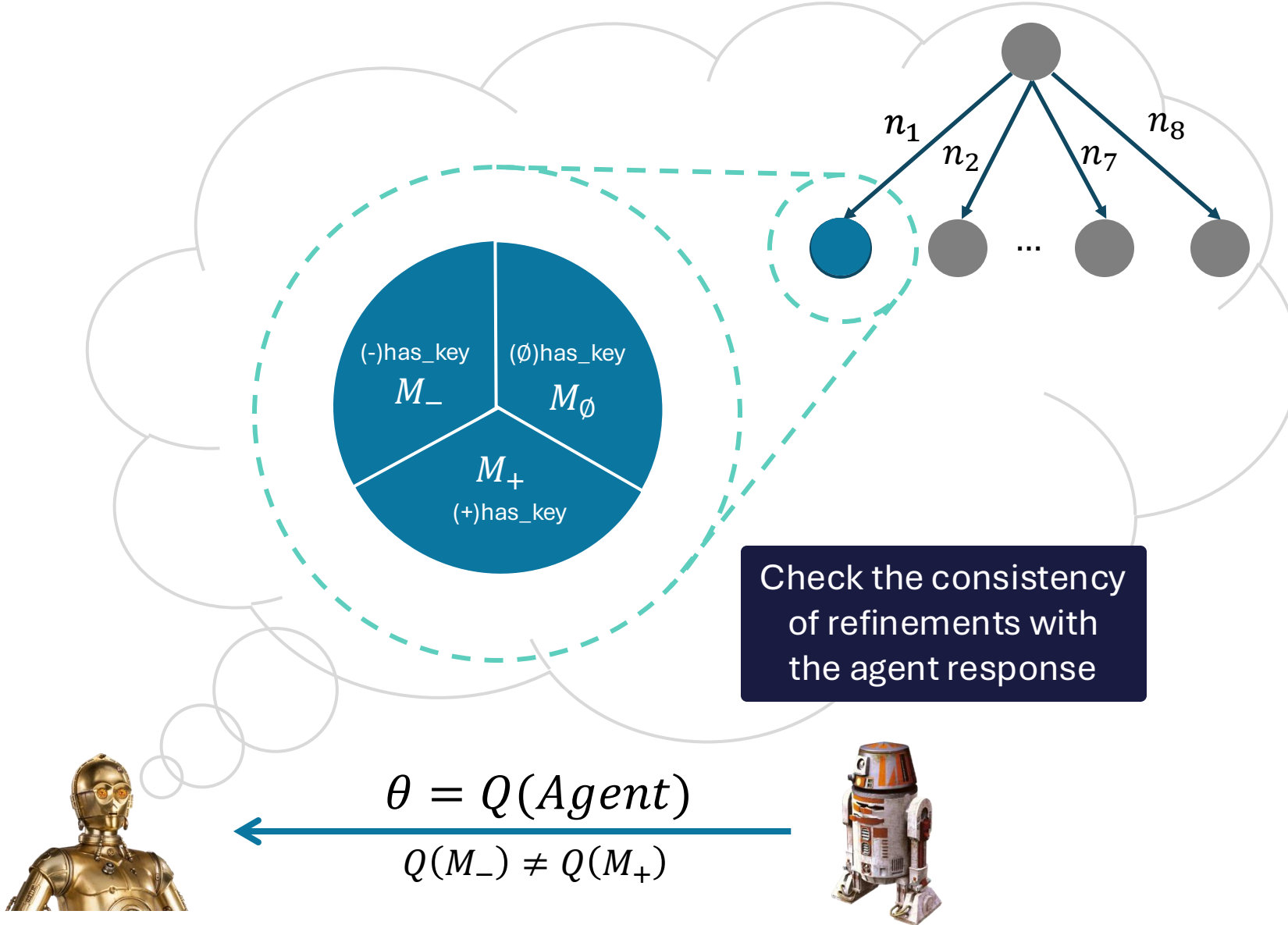
```
(:action open-door
  :parameters (?l1)
  :precondition (and
    n1 (+/-/∅)(has_key)
    n2 (+/-/∅)(door_open)
    n3 (+/-/∅)(door_adjacent ?l1)
    n4 (+/-/∅)(player_at ?l1))
  :effect (and
    n5 (+/-/∅)(has_key)
    n6 (+/-/∅)(door_open)
    n7 (+/-/∅)(door_adjacent ?l1)
    n8 (+/-/∅)(player_at ?l1)))
```

Hierarchical Query Synthesis



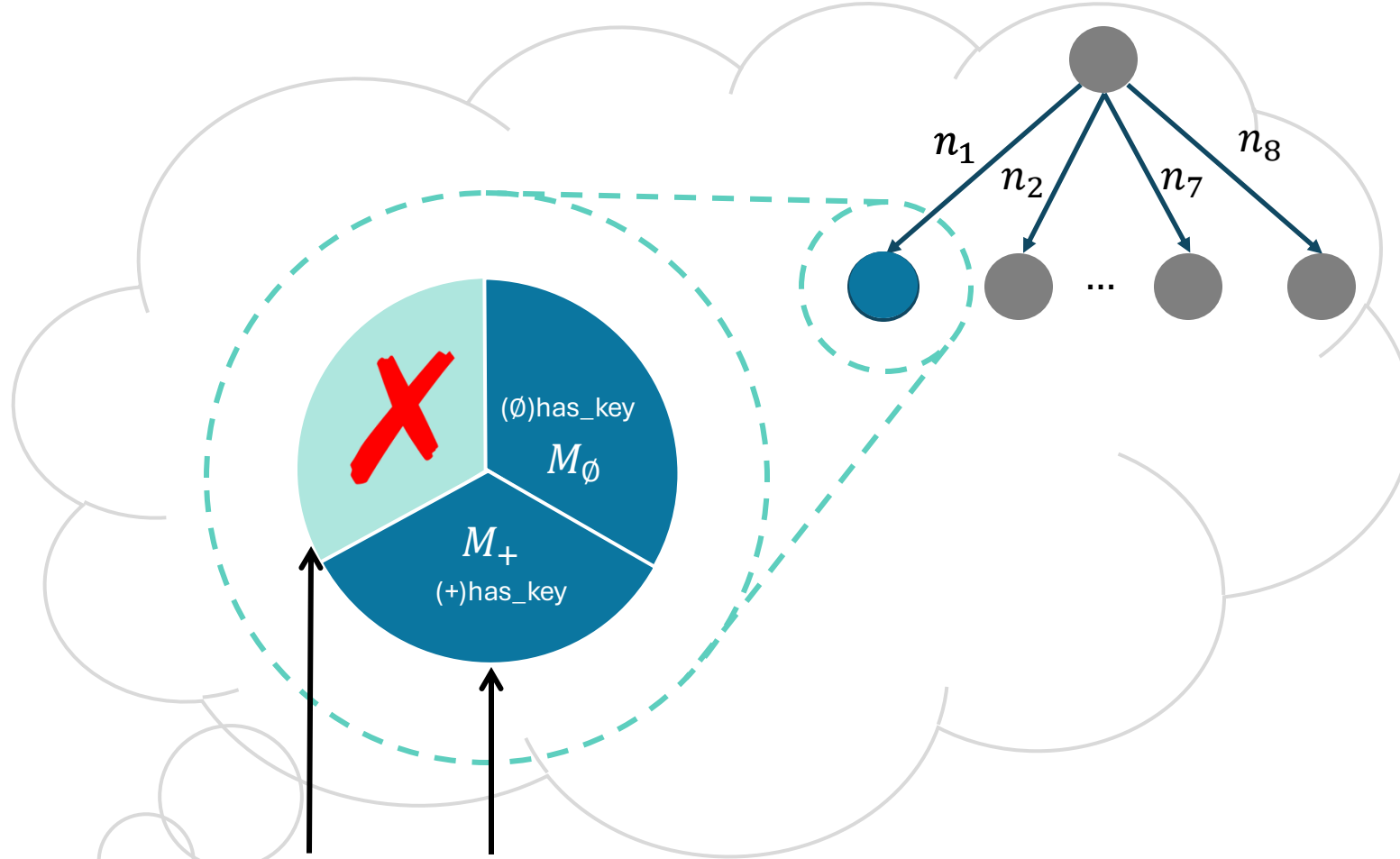
```
(:action open-door
  :parameters (?l1)
  :precondition (and
     $n_1$   $(+/-/\emptyset)(\text{has\_key})$ 
     $n_2$   $(+/-/\emptyset)(\text{door\_open})$ 
     $n_3$   $(+/-/\emptyset)(\text{door\_adjacent } ?l1)$ 
     $n_4$   $(+/-/\emptyset)(\text{player\_at } ?l1)$ 
  )
  :effect (and
     $n_5$   $(+/-/\emptyset)(\text{has\_key})$ 
     $n_6$   $(+/-/\emptyset)(\text{door\_open})$ 
     $n_7$   $(+/-/\emptyset)(\text{door\_adjacent } ?l1)$ 
     $n_8$   $(+/-/\emptyset)(\text{player\_at } ?l1)$ 
  ))
```

Hierarchical Query Synthesis



```
(:action open-door
:parameters (?l1)
:precondition (and
  n1 (+/-/∅)(has_key)
  n2 (+/-/∅)(door_open)
  n3 (+/-/∅)(door_adjacent ?l1)
  n4 (+/-/∅)(player_at ?l1))
:effect (and
  n5 (+/-/∅)(has_key)
  n6 (+/-/∅)(door_open)
  n7 (+/-/∅)(door_adjacent ?l1)
  n8 (+/-/∅)(player_at ?l1))
```

Hierarchical Query Synthesis

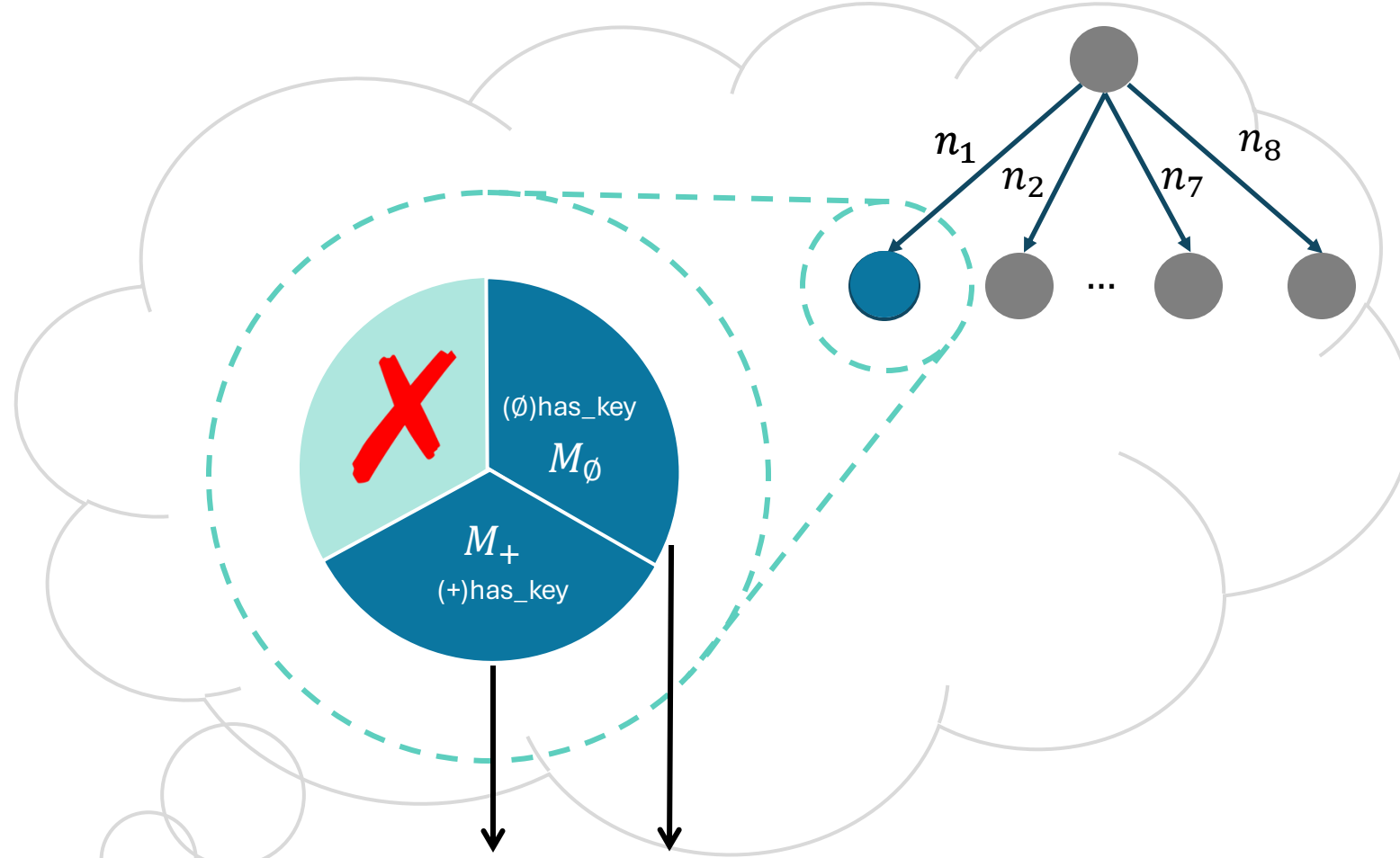


```
(:action open-door
  :parameters (?l1)
  :precondition (and
    n1 (+/∅)(has_key)
    n2 (+/-/∅)(door_open)
    n3 (+/-/∅)(door_adjacent ?l1)
    n4 (+/-/∅)(player_at ?l1))
  :effect (and
    n5 (+/-/∅)(has_key)
    n6 (+/-/∅)(door_open)
    n7 (+/-/∅)(door_adjacent ?l1)
    n8 (+/-/∅)(player_at ?l1))
```



Reject abstract model(s) that are not consistent with the agent

Hierarchical Query Synthesis

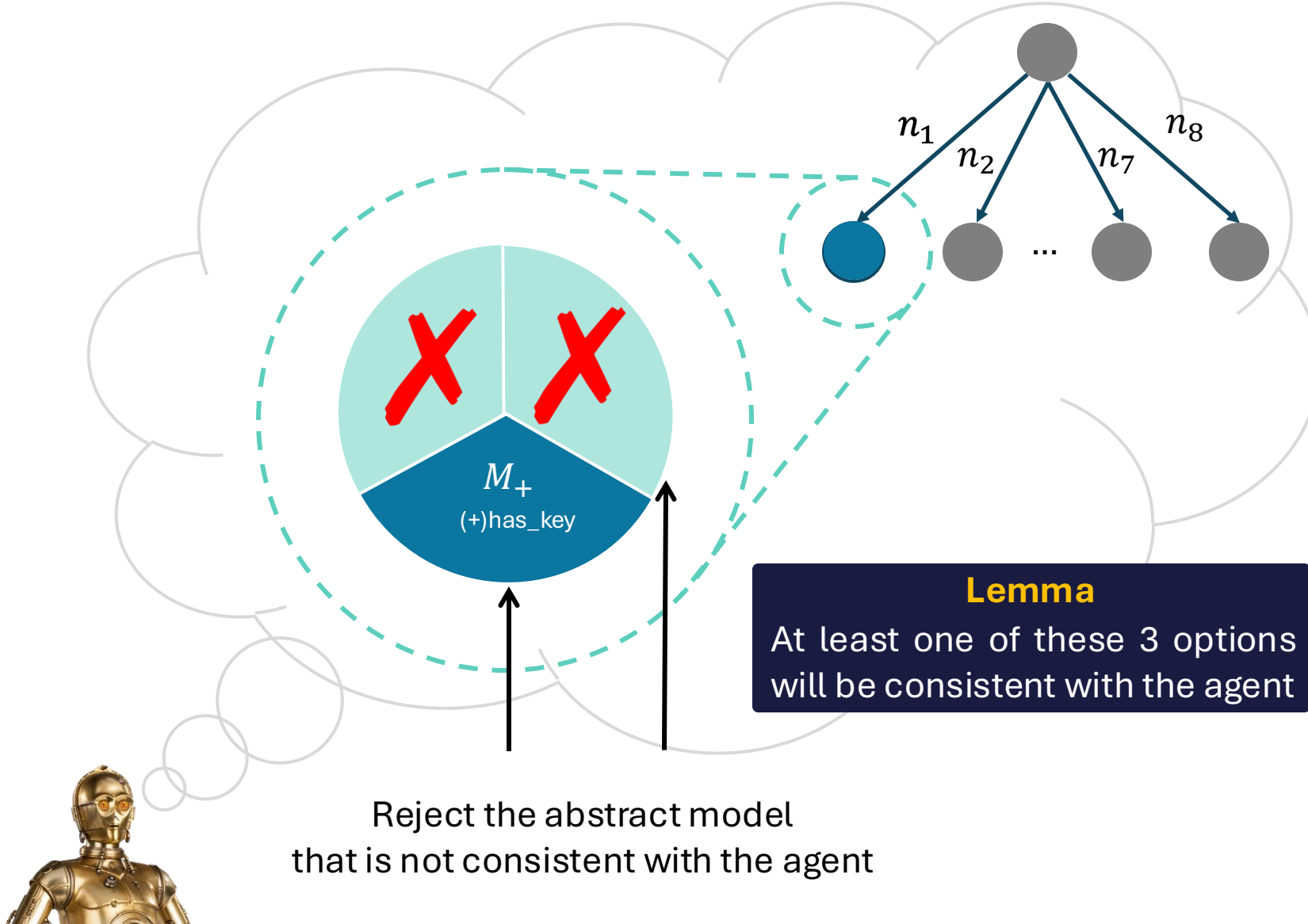


Generate a distinguishing query
for these two abstract models

```
(:action open-door
  :parameters (?l1)
  :precondition (and
    n1 (+/∅)(has_key)
    n2 (+/-/∅)(door_open)
    n3 (+/-/∅)(door_adjacent ?l1)
    n4 (+/-/∅)(player_at ?l1))
  :effect (and
    n5 (+/-/∅)(has_key)
    n6 (+/-/∅)(door_open)
    n7 (+/-/∅)(door_adjacent ?l1)
    n8 (+/-/∅)(player_at ?l1))
```

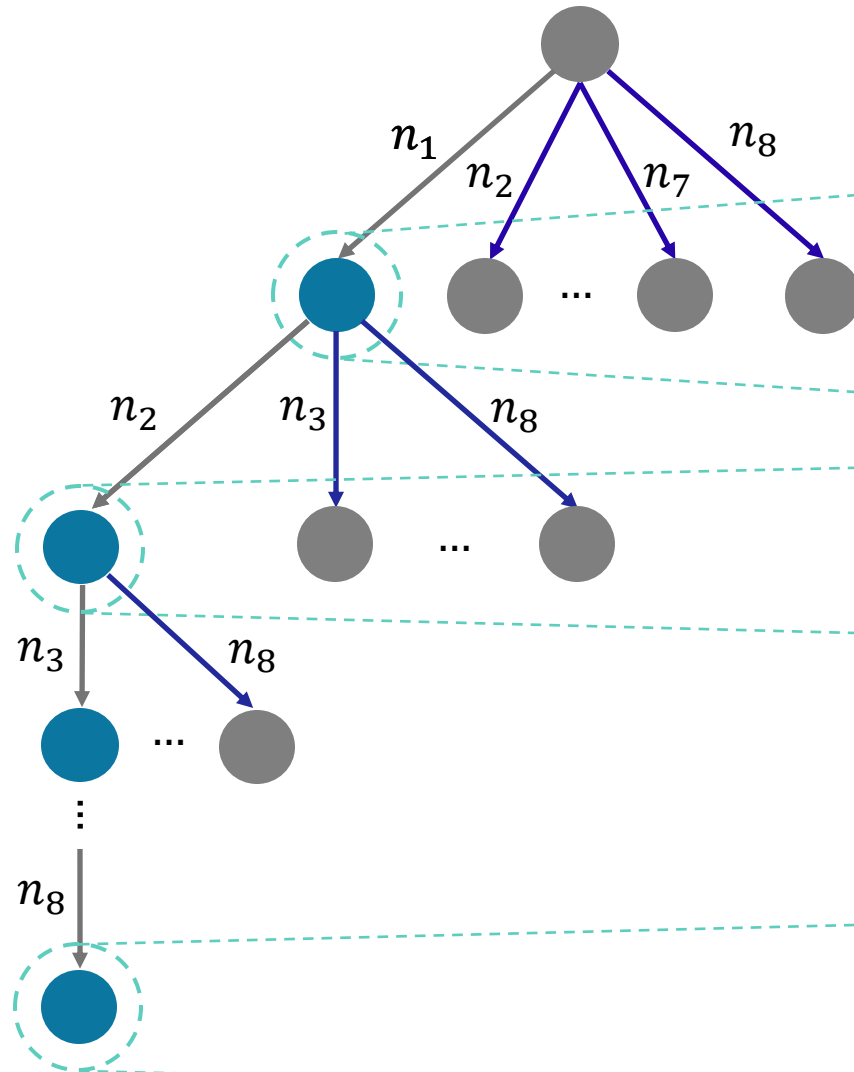


Hierarchical Query Synthesis



```
(:action open-door
:parameters (?l1)
:precondition (and
n1 (+) (has_key)
n2 (+/-/∅) (door_open)
n3 (+/-/∅) (door_adjacent ?l1)
n4 (+/-/∅) (player_at ?l1))
:effect (and
n5 (+/-/∅) (has_key)
n6 (+/-/∅) (door_open)
n7 (+/-/∅) (door_adjacent ?l1)
n8 (+/-/∅) (player_at ?l1))
```

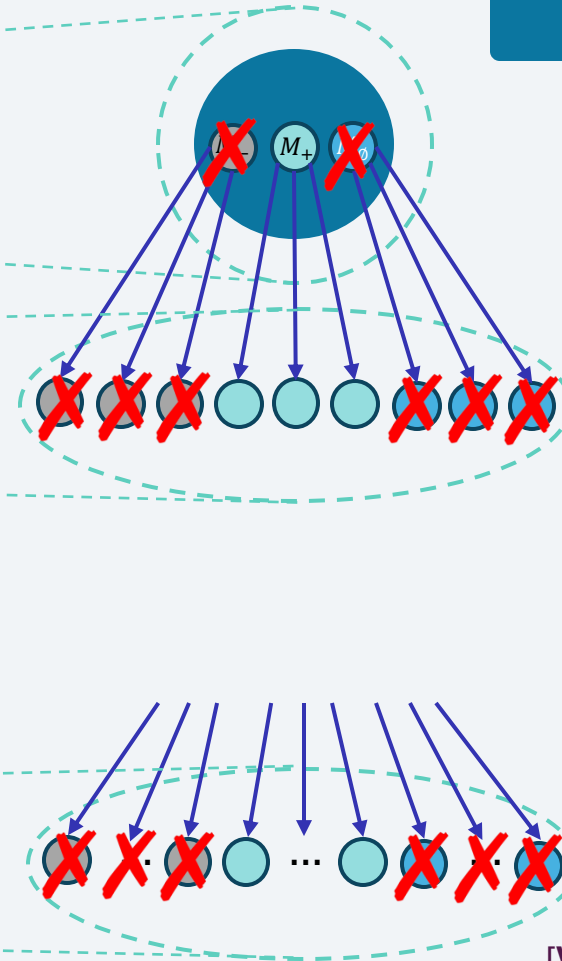
Hierarchical Query Synthesis



Key feature of the algorithm

Whenever we prune an abstract model, we prune a large number of concrete models.

Active Learning



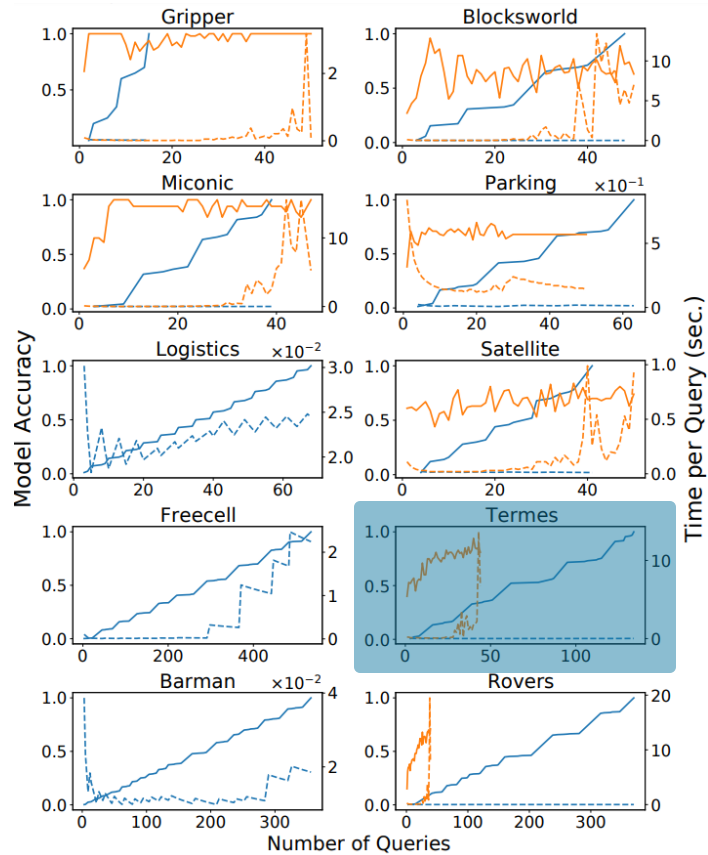
AAM learns Accurate Model with fewer Queries

- Asses by learning the model and compare with ground truth.
- Baseline[†]: A passive learner (FAMA) that observes agent behavior



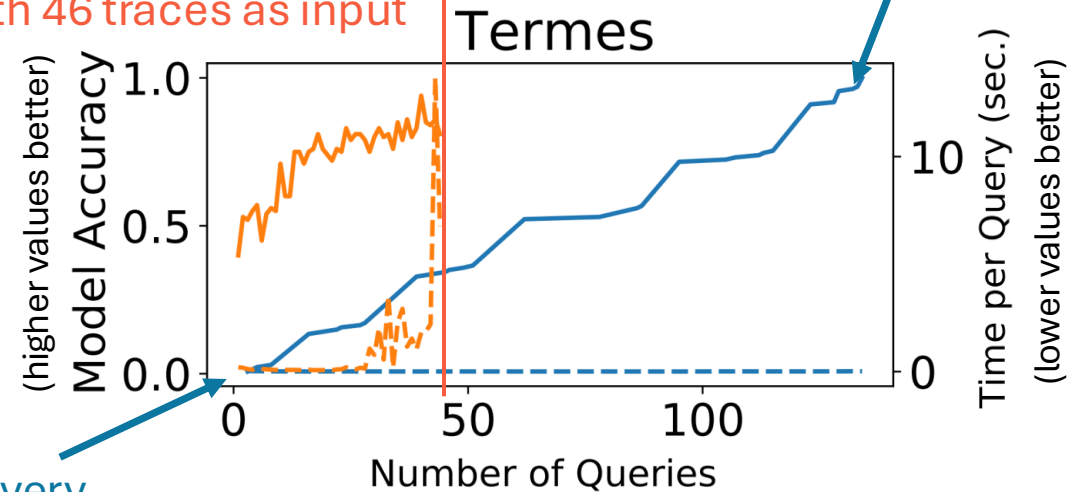
Random deterministic planning agent from IPC

Accuracy: — AAM — FAMA
Time: - - - AAM - - - FAMA



FAMA ran out of memory with 46 traces as input

AAM learned the correct model with 134 queries



AAM takes very less time

[Verma, Marpally, Srivastava; AAAI '21]

AAM learns Accurate Deterministic Models

- Theorem (*termination*) : The algorithm terminates after a finite number of iterations.
- Theorem (*soundness*): The resulting (set of) model(s) is(are) functionally equivalent to the ground truth model.

Theorem 3 in Paper



Stochastic and Stationary Setting

Input

- Predicates (User vocabulary)
 - With their evaluation functions
- List of capabilities.

Output

- PPDDL-like description of each capability.

Assumptions

- User's vocabulary matches simulator's vocabulary.
- Black-Box AI provides a list of capabilities.
- Stationary agent model.
- ~~Deterministic~~ ^{Stochastic} environment.
- Fully observable setting.

Discovering Capabilities

Input

- Predicates (User vocabulary)
 - With their evaluation functions
- Samplers: high-level state to low-level state.
- Low-level state transitions.



Output

- List of capabilities.
- PDDL-like description of each capability.

[Verma, Marpally, Srivastava; KR '22]

Assumptions

- ~~User's vocabulary matches simulator's vocabulary.~~
- Black-Box AI provides a list of ~~capabilities.~~ **transitions.**
- Stationary agent model.
- Deterministic environment.
- Fully observable setting.

Differential Assessment

Input

- Initial model of the AI system.
 - Predicates (User vocabulary)
 - With their evaluation functions
 - List of capabilities.
- Observations of AI system working in the environment.

Output

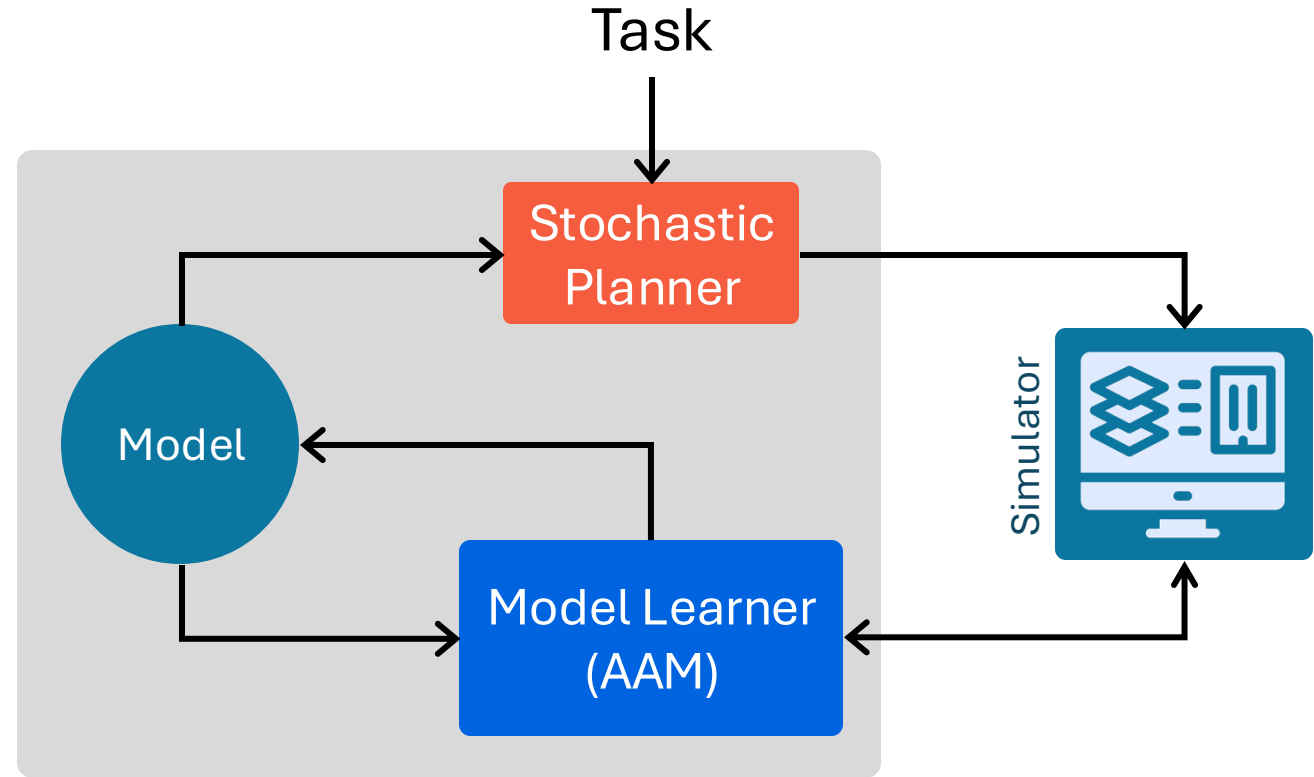
- Updated PDDL-like description of each capability.

Assumptions

- User's vocabulary matches simulator's vocabulary.
- Black-Box AI provides a list of capabilities.
- ~~Stationary~~ ^{Adaptive} agent model.
- Deterministic environment.
- Fully observable setting.

Continual Learning and Planning (CLaP)

- RL setting with stream of input tasks.
- Maximize #tasks completed within a fixed budget.
- Need to adapt fast (minimize adaptive delay), and compute good solutions (minimize regret).



01

Research Area &
Interesting Problems

02

Agent Interrogation
Algorithm
(AAAI 2021)

03

Future Research
Directions

04

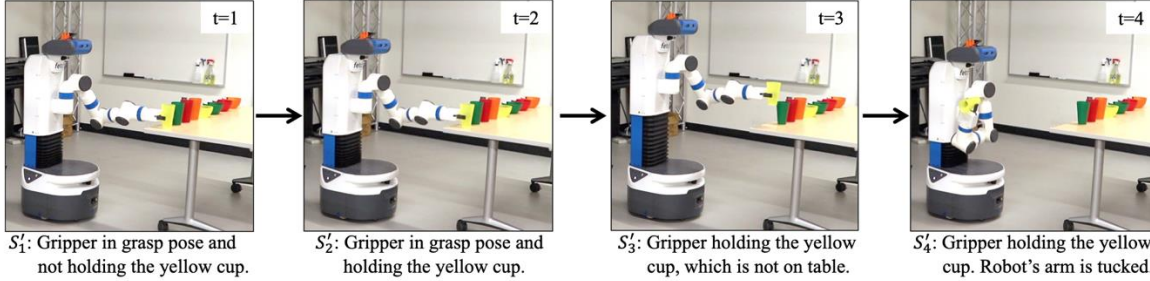
Publications

05

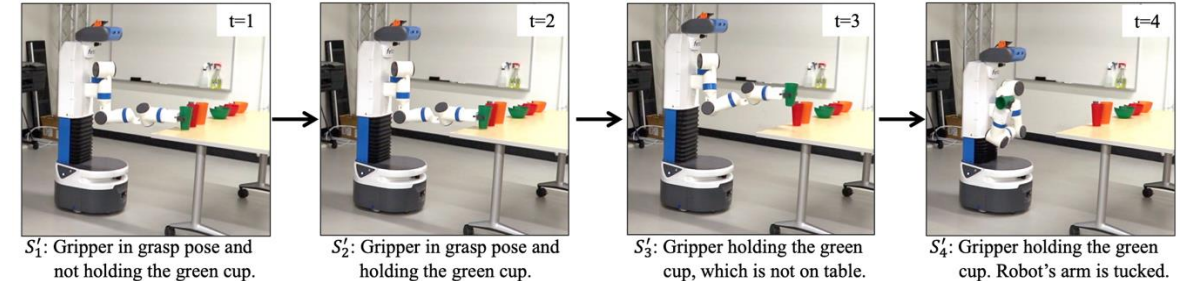
Long Term
Research Plan

Autonomous Robot Skill Acquisition

Yellow Cup



Green Cup



- $S_1 : \{(table-can1 \ ?table \ ?can), (can-gripper1 \ ?can \ ?gripper), (base-gripper0 \ ?base \ ?gripper), (base-table1 \ ?base \ ?table)\}$.
- $S_2 : \{(table-can1 \ ?table \ ?can), (can-gripper2 \ ?can \ ?gripper), (base-gripper0 \ ?base \ ?gripper), (base-table1 \ ?base \ ?table)\}$.
- $S_3 : \{(table-can0 \ ?table \ ?can), (can-gripper2 \ ?can \ ?gripper), (base-gripper0 \ ?base \ ?gripper), (base-table1 \ ?base \ ?table)\}$.
- $S_4 : \{(table-can0 \ ?table \ ?can), (can-gripper2 \ ?can \ ?gripper), (base-gripper1 \ ?base \ ?gripper), (base-table1 \ ?base \ ?table)\}$.

First-order State

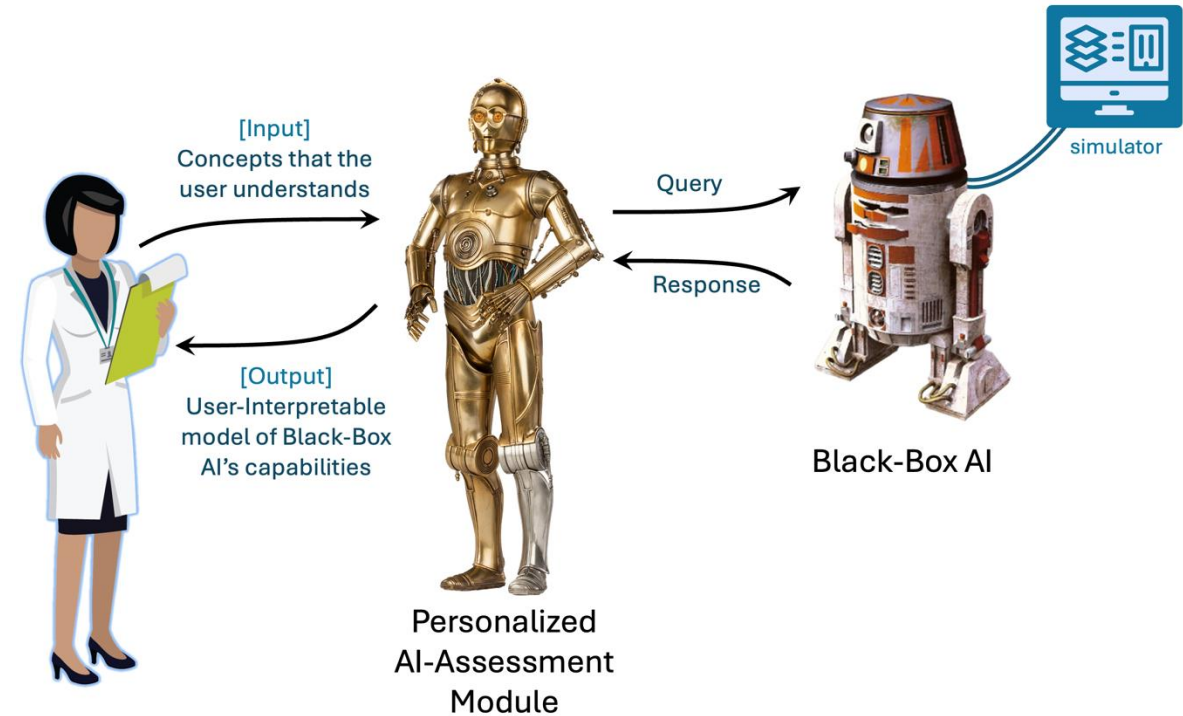
- $C_{12} = \langle +C_{12} = \{(can-gripper2 \ ?can \ ?gripper)\}, -C_{12} = \{(can-gripper1 \ ?can \ ?gripper)\} \rangle$.
- $C_{23} = \langle +C_{23} = \{(table-can0 \ ?table \ ?can)\}, -C_{23} = \{(table-can1 \ ?table \ ?can)\} \rangle$.
- $C_{34} = \langle +C_{34} = \{(base-gripper1 \ ?base \ ?gripper)\}, -C_{34} = \{(base-gripper0 \ ?base \ ?gripper)\} \rangle$.

Learned Skills

- Currently* depends on manual selection of tasks.
- Goal is to make the task selection autonomous.
 - Using active learning techniques.

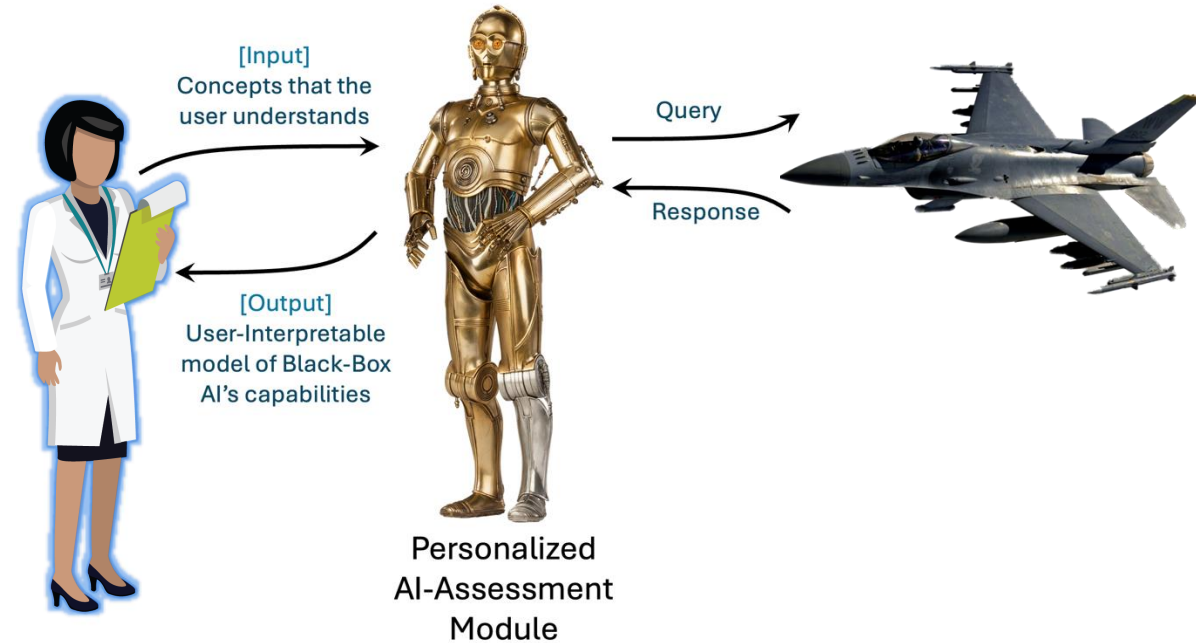
Interpretable Representations

- Representations beyond PDDL.
 - LTL – Linear Temporal Logic
 - STL- Signal Temporal Logic
 - CDL- Crow Definition Language
- How interpretable each representation is?



Real World Setting

- Flying Extended Trail
- Human must collaborate with the AI-powered aircraft.
- How can we explain the aircraft's policy to the human in an interpretable way?



01

Research Area &
Interesting Problems

02

Agent Interrogation
Algorithm
(AAAI 2021)

03

Future Research
Directions

04

Publications

05

Long Term
Research Plan

Contributions

AI Assessment

[**Verma**, Marpally, Srivastava; AAAI '21]

[Nayyar*, **Verma***, Srivastava; AAAI '22]

[**Verma**, Marpally, Srivastava; KR '22]

[**Verma**, Karia, Srivastava; NeurIPS '23]

[**Verma***, Karia*, Vipat, Gupta, Srivastava; GenPlan '23]

[**Verma**, Srivastava; AAAI '24 SS]

[Karia, Dobhal, Bramblett, **Verma**, Srivastava; AAAI '24 SS]

Generalization in Planning

[Karia*, **Verma***, Speranzon, Srivastava; ICAPS '24]

[Shah, Nagpal, **Verma**, Srivastava; Preprint]

Causal Accuracy of Symbolic Models

[**Verma**, Srivastava; GenPlan '21]

[**Verma**, Srivastava; Tech Report, 2021]

Explainable Robot Planning

[Shah*, **Verma***, Angle, Srivastava; AAMAS '22]

[Dobhal, Nagpal, Karia, **Verma**, Nayyar, Shah, Srivastava; In Submission]

[Dadvar, Majd, Oikonomou, **Verma**, Fainekos, Srivastava; In Submission]

Contributions

Conferences

- **Pulkit Verma**, Shashank Rao Marpally, and Siddharth Srivastava. *Asking the Right Questions: Learning Interpretable Action Models through Query Answering*. In AAAI 2021.
 - Rashmeet Kaur Nayyar*, **Pulkit Verma***, and Siddharth Srivastava. *Asking the Right Questions: Learning Interpretable Action Models through Query Answering*. In AAAI 2022.
 - **Pulkit Verma**, Shashank Rao Marpally, and Siddharth Srivastava. *Discovering User-Interpretable Capabilities of Black-Box Planning Agents*. In KR 2022.
 - Naman Shah*, **Pulkit Verma***, Trevor Angle, and Siddharth Srivastava. *JEDAI: A System for Skill-Aligned Explainable Robot Planning*. In AAMAS 2022 (Demonstration Track).
- A blue-bordered box containing a gold trophy icon on the left and the text "Winner of Best Demo Award" in a blue serif font on the right.
- Yizhong Wang et al. *Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ Tasks*. In EMNLP 2022.
 - **Pulkit Verma**, Rushang Karia, and Siddharth Srivastava. *Autonomous Capability Assessment of Sequential Decision-Making Systems in Stochastic Settings*. In NeurIPS 2023.
 - Rushang Karia*, **Pulkit Verma***, Alberto Speranzon, and Siddharth Srivastava. *Epistemic Exploration for Generalizable Planning and Learning in Non-Stationary Settings*. In ICAPS 2024.

Workshops, Symposia, Preprints

- **Pulkit Verma** and Siddharth Srivastava. *Learning Generalized Models by Interrogating Black-Box Autonomous Agents*. In AAAI 2020 GenPlan.
- **Pulkit Verma** and Siddharth Srivastava. *Learning Causal Models of Autonomous Agents using Interventions*. In IJCAI 2021 GenPlan.
- **Pulkit Verma**, Shashank Rao Marpally, and Siddharth Srivastava. *Learning User-Interpretable Descriptions of Black-Box AI System Capabilities*. In ICAPS 2021 KEPS.
- **Pulkit Verma***, Rushang Karia*, Gaurav Vipat, Anmol Gupta, and Siddharth Srivastava. *Learning AI-System Capabilities under Stochasticity*. In NeurIPS 2023 GenPlan.
- **Pulkit Verma** and Siddharth Srivastava. *User-Aligned Autonomous Capability Assessment of Black-Box AI Systems*. In AIA 2024.
- Rushang Karia, Daksh Dobhal, Daniel Bramblett, **Pulkit Verma**, and Siddharth Srivastava. *Can LLMs translate SATisfactorily? Assessing LLMs in Generating and Interpreting Formal Specifications*. In AIA 2024.
- Mehdi Dadvar, Keyvan Majd, Elena Oikonomou, **Pulkit Verma**, Georgios Fainekos, and Siddharth Srivastava. *Joint Communication and Motion Planning for Cobots in Real-World Contexts*. In Submission.
- Daksh Dobhal, Jayesh Nagpal, Rushang Karia, **Pulkit Verma**, Rashmeet Kaur Nayyar, Naman Shah, and Siddharth Srivastava. *Using Explainable AI and Hierarchical Planning for Outreach with Robots*. In Submission.
- Naman Shah, Jayesh Nagpal, **Pulkit Verma**, and Siddharth Srivastava. *From Reals to Logic and Back: Inventing Symbolic Vocabularies, Actions, and Models for Planning from Raw Data*. Preprint.

*Equal Contribution

01

Research Area &
Interesting Problems

02

Agent Interrogation
Algorithm
(AAAI 2021)

03

Future Research
Directions

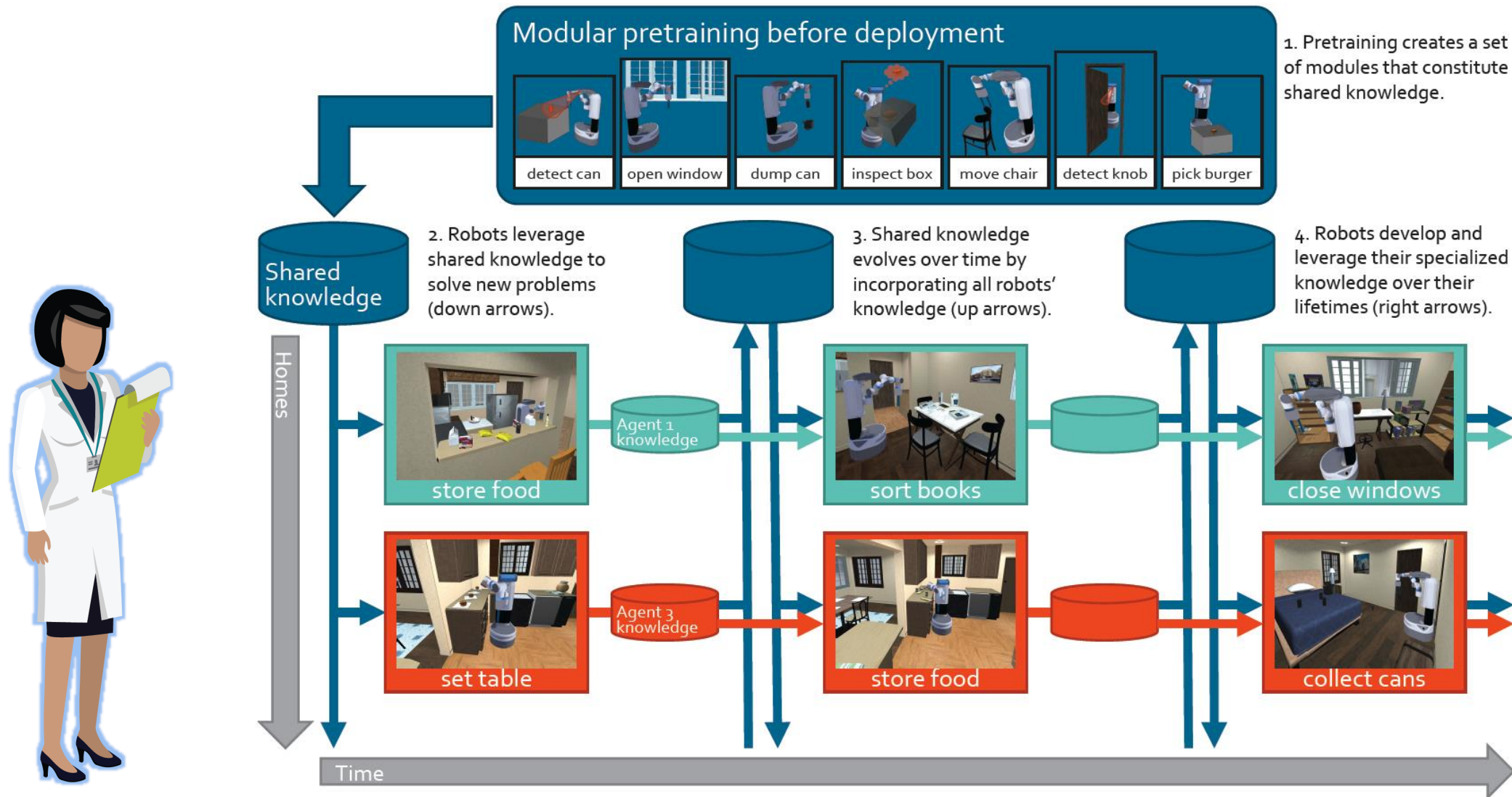
04

Publications

05

Long Term
Research Plan

Lifelong Learning with Human-in-the Loop



Warehouse Robotics

- The increasing number of Indian robotics companies makes research accessible with cheaper robots.
- Possible work on removing the need for warehouse robots to follow fixed paths.
- Can start research on Future of work where humans and robots work in same space.

ADDVERB
Human | Robot | Possibilities

Noida-based
Warehouse Robot Manufacturer



DYNAMO 100 KG



DYNAMO 1000 KG



DYNAMO 200 KG



DYNAMO 500 KG



Dynamo 2500



ZIPPY 10 KG



ZIPPY 100 KG



ZIPPY 25 KG



Rushang Karia



Shashank Marpally



Rashmeet Nayyar



Alberto Speranzon

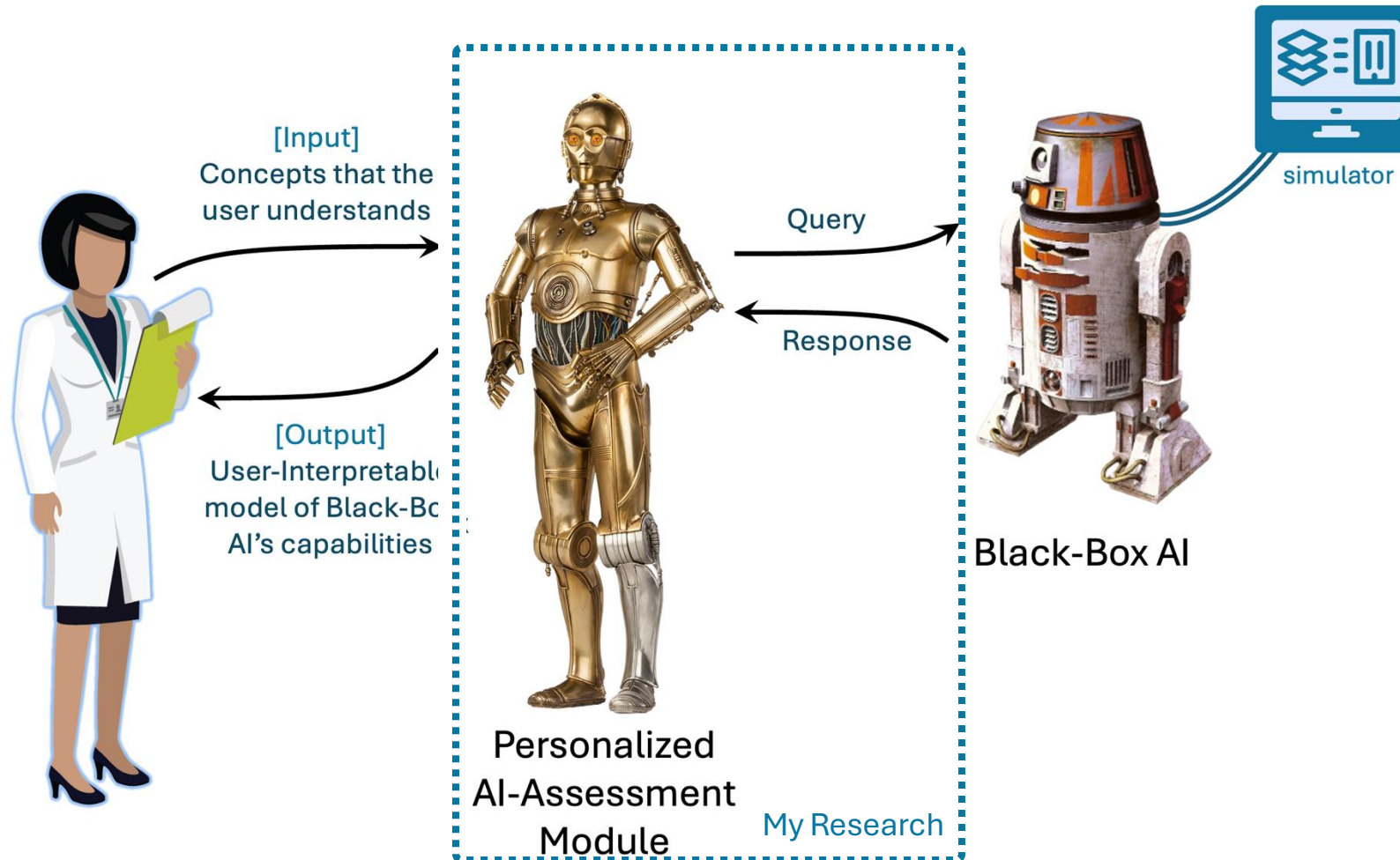


Siddharth Srivastava



Learning Interpretable Models for Personalized Assessment of Taskable AI Systems

Pulkit Verma
pulkitv@mit.edu



Questions?

bit.ly/pulkit-iitrpr



bitly