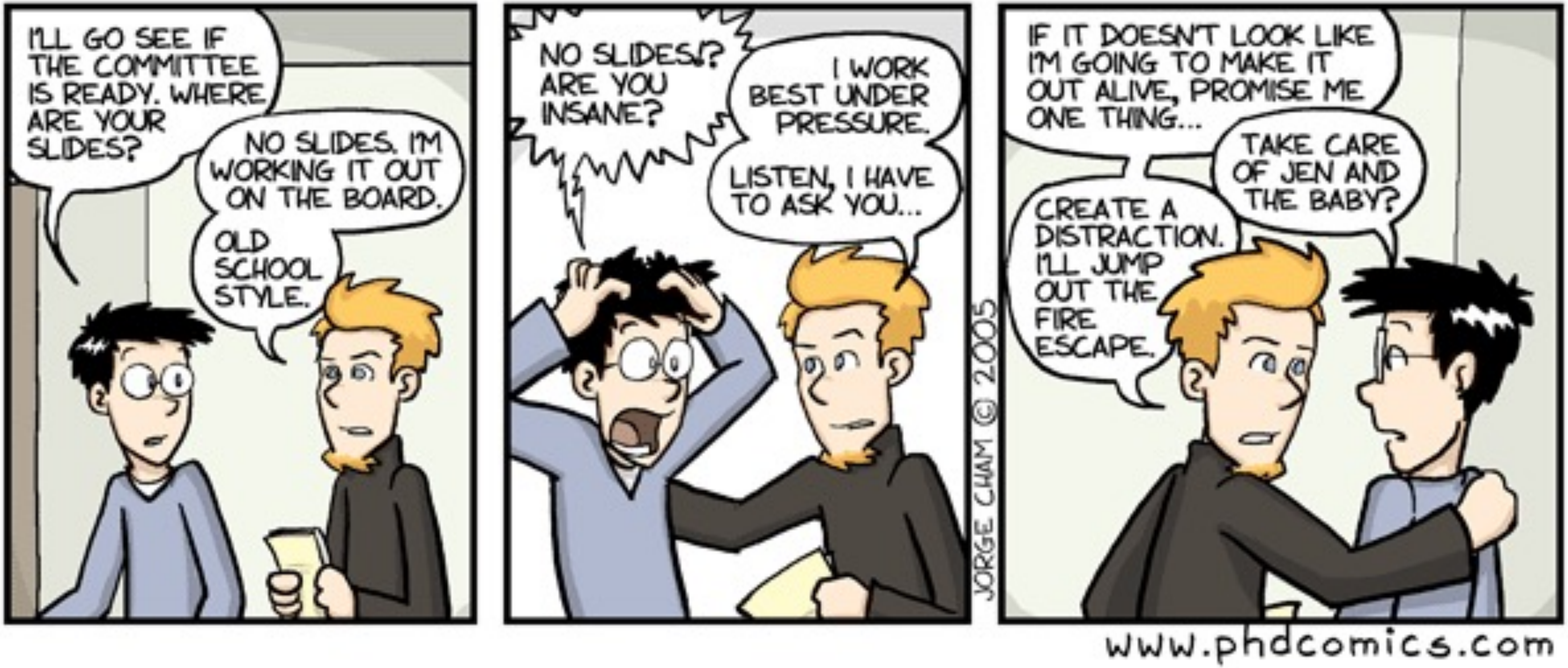

I'LL GO SEE IF THE COMMITTEE IS READY. WHERE ARE YOUR SLIDES?
NO SLIDES. I'M WORKING IT OUT ON THE BOARD.
OLD SCHOOL STYLE.
NO SLIDES!? ARE YOU INSANE?
I WORK BEST UNDER PRESSURE.
LISTEN, I HAVE TO ASK YOU...
IF IT DOESN'T LOOK LIKE I'M GOING TO MAKE IT OUT ALIVE, PROMISE ME ONE THING...
TAKE CARE OF JEN AND THE BABY?
CREATE A DISTRACTION. I'LL JUMP OUT THE FIRE ESCAPE.
JORGE CHAM © 2005
www.phdcomics.com

PhD Dissertation Defense

April 03, 2024

Data-Efficient Paradigms for Personalized Assessment of Taskable AI Systems

Pulkit Verma

verma.pulkit@asu.edu

pulkitverma.net

A A I R
Autonomous Agents and Intelligent Robots

Arizona State University

Dissertation Committee

Siddharth Srivastava [Chair]

Nancy Cooke

Georgios Fainekos

Yu Zhang

01
Introduction

02
Foundational Approach

03
Generalizations

04
Applications

05
Conclusion

Data-Efficient Paradigms for

Personalized Assessment of

Adaptive Black-Box

Taskable AI Systems

Taskable AI Systems: Systems that can Learn and Plan

User gives a task and Robots have to complete it

- Sequential Decision-Making Systems.
- Systems designed to be able to help user.
- User has a task in mind and expects the AI system to help in that task.

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?

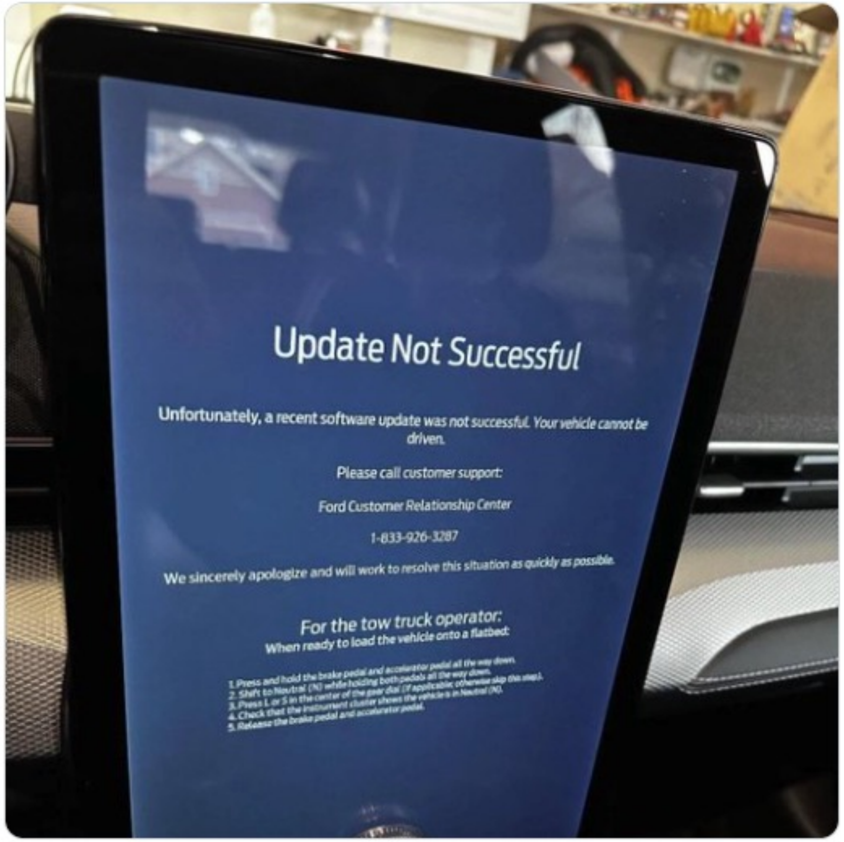
- AI systems should support third-party assessment.

- The assessment should work with:
  - Adaptive AI systems.
  - Black-Box AI systems.

Adaptive Taskable AI Systems

Dan Luu
@danluu · Follow

"Unfortunately, a recent software update was not successful. Your vehicle cannot be driven. Please call customer support:"



1:46 PM · Dec 25, 2023

USA TODAY

Tesla self-driving software update begins rollout though company says to use with caution

Charisse Jones, USA TODAY
July 12, 2021 · 2 min read

Lucid Owners Facing Software Glitches That Brick EVs Or Drive the Wrong Direction

1

Owen Bellwood
November 10, 2022 · 2 min read

AEG combi microwave thinks it is a steam oven and no longer works after an incorrect update

By Julian Huijbregts 04-03-2022 • 10:48
News editor

NEWS

Nest thermostat software bug chills users once again

A faulty software update for the smart thermostat made batteries drain and home temperatures drop.

By Jared Newman
TechHive | JAN 14, 2016 8:38 AM PST

Desiderata of Assessment System

Interpretability	Correctness
Generalizability	Easy to Satisfy Requirements

The Assessment Problem

Will it be able to safely rearrange my lab for the next round of experiments?

Output

- A description of agent's working:
  - A list of capabilities.
  - An interpretable description of each capability.

Black-Box AI

- Arbitrary internal implementation
- Comes with a Trained Policy

Related Work

Improving Agent Capabilities

Better Sequential Decision-Making Algorithms

E.g., Learning High-Level Actions, etc. ……

Konidaris et al. (JAIR'18)
James et al. (ICML'20)
Allen et al. (IJCAI'21)

Learning Interpretable Descriptions of Capabilities of a Given Agent

Learning Functionality from Passive Observations

Stern et al. (IJCAI'17)
Cresswell et al. (ICAPS'09)
Yang et al. (AIJ 2007)
Zhuo et al. (IJCAI'13)
Aineto et al. (AIJ'19)

Discovering and Learning Agent Capabilities

[Our Approach]

[Input]

Learner

Description of the AI system's working

[Output]

- No other work on assessment.
- The closest work is on learning from passive observations.
- Learn what the agent does, not what it can do when it is retasked.

OBD Scanner for Black-Box AI?

Can we build something like an OBD scanner for Black-Box AI Systems?

Yes!!

But we'll need something more general and powerful.

Cars have:

- Well-understood components
- Known internal design
- (Commonly) limited functionality
- Not as versatile as a household robot

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

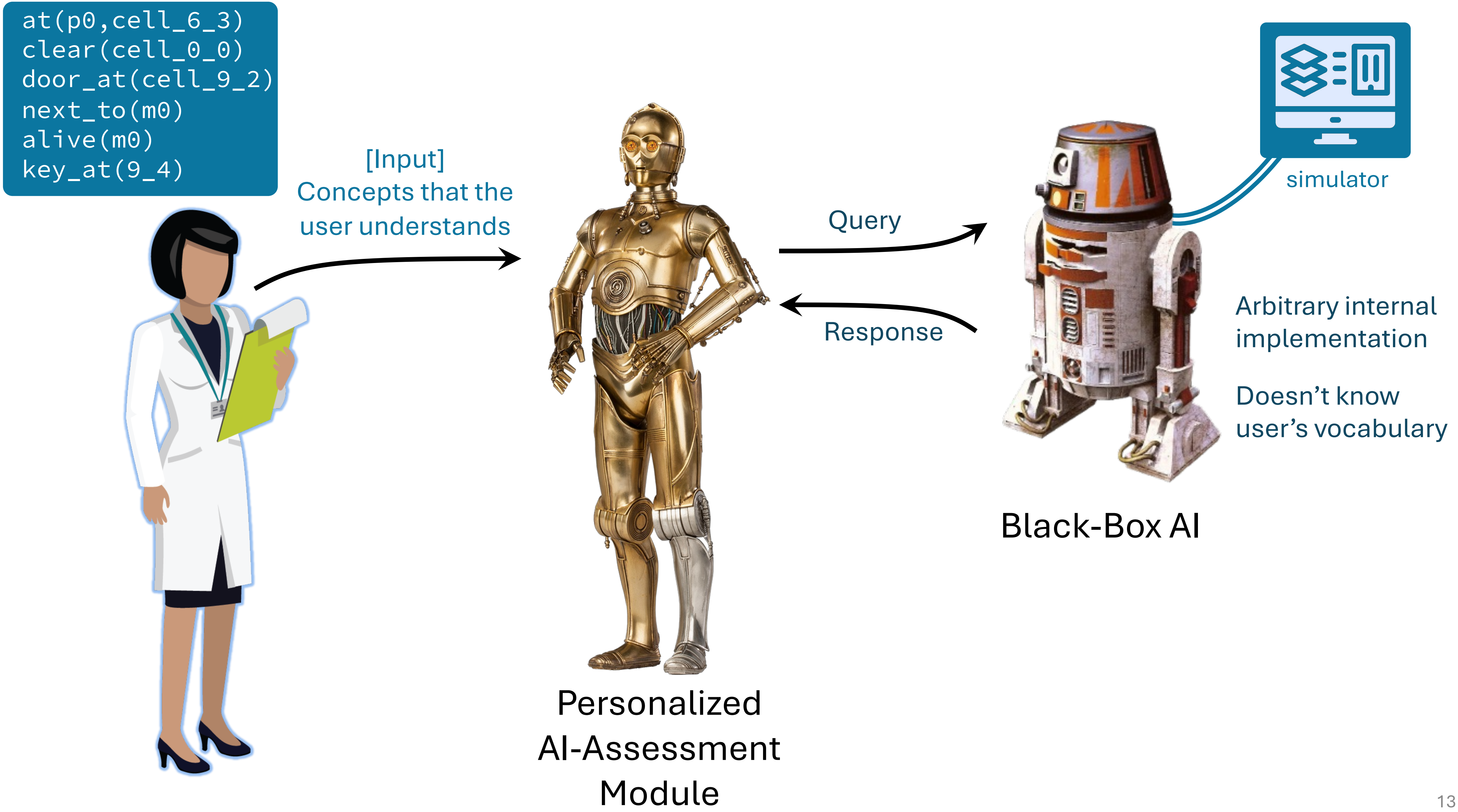
[Input]
Concepts that the user understands

Personalized AI-Assessment Module (AAM)

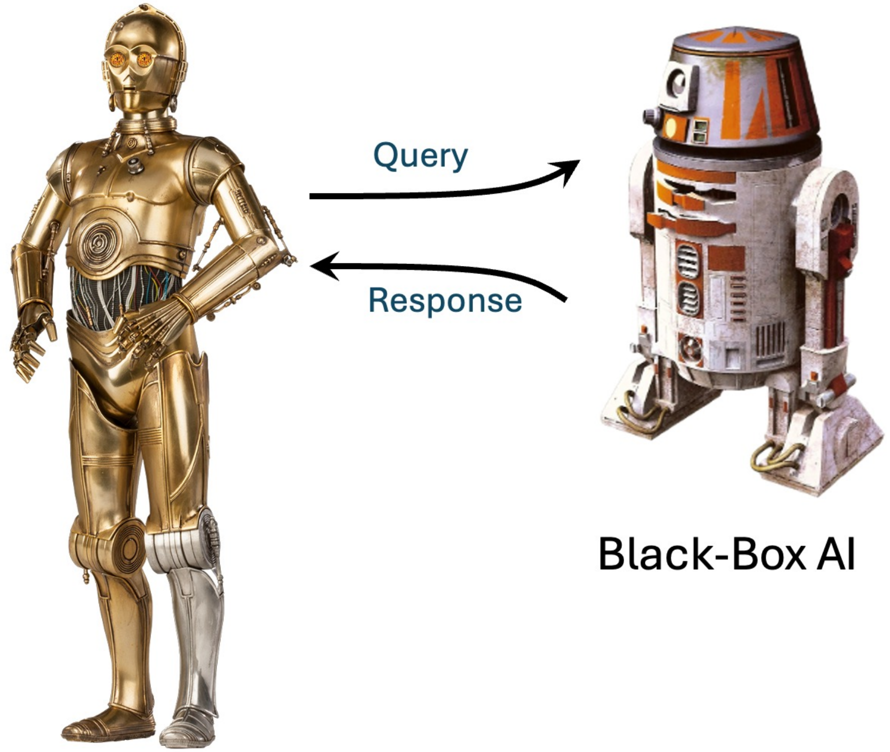
Arbitrary internal implementation

Doesn't know user's vocabulary

Black-Box AI

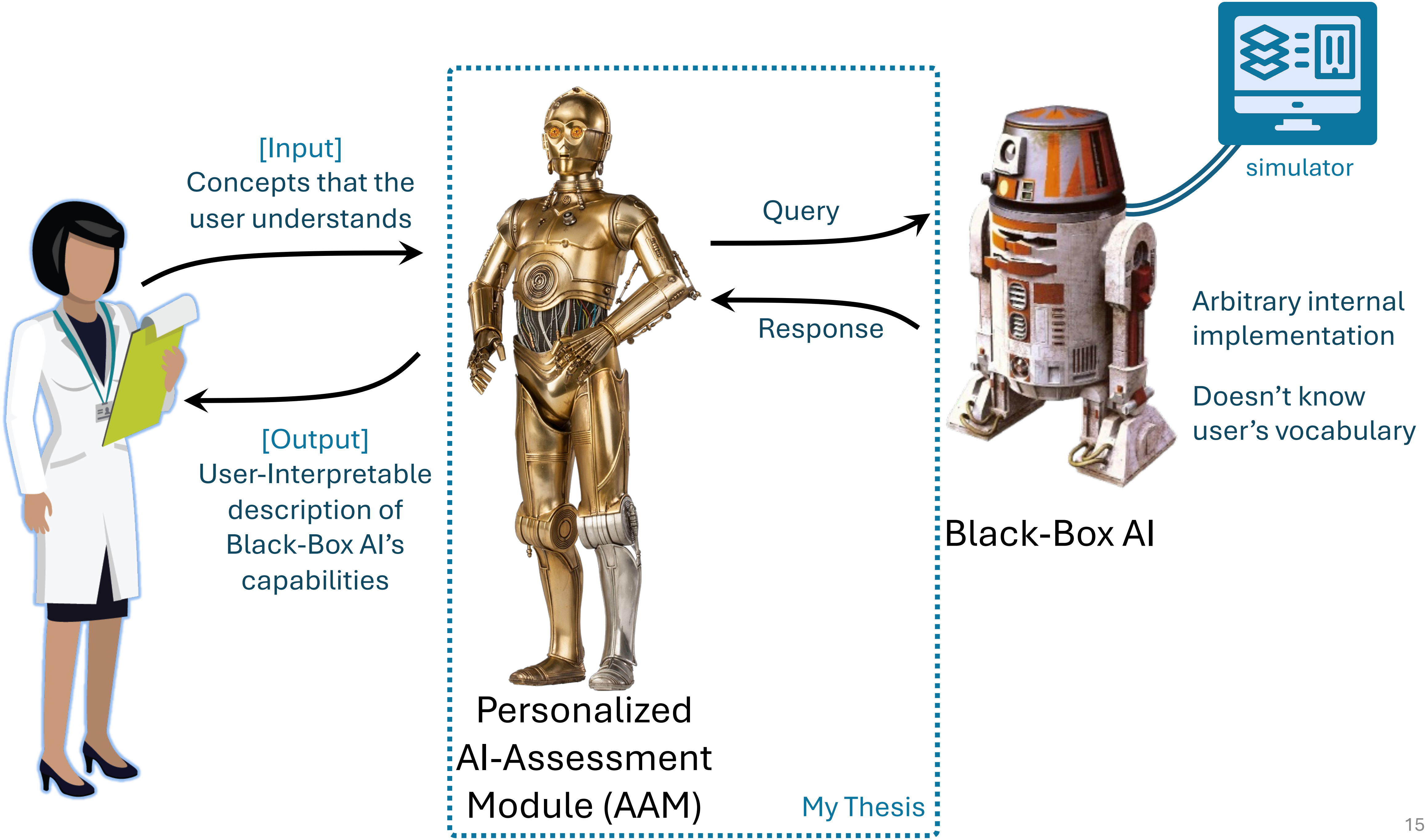

at(p0,cell_6_3)
clear(cell_0_0)
door_at(cell_9_2)
next_to(m0)
alive(m0)
key_at(9_4)
[Input]
Concepts that the user understands
Query
Response
simulator
Arbitrary internal implementation
Doesn’t know user’s vocabulary
Black-Box AI
Personalized AI-Assessment Module

Black-Box AI System Interface



Personalized
AI-Assessment
Module

- Simple Query-Response Interface
- Should work for a variety of taskable AI systems.
    - Independent of their internals.
- Requirement: $\langle$QueryType, ResponseType, $\rho\rangle$.


[Input]
Concepts that the user understands
[Output]
User-Interpretable description of Black-Box AI's capabilities
Query
Response
Personalized AI-Assessment Module (AAM)
My Thesis
simulator
Arbitrary internal implementation
Doesn't know user's vocabulary
Black-Box AI

The Assessment Problem

Black-Box (Taskable) AI

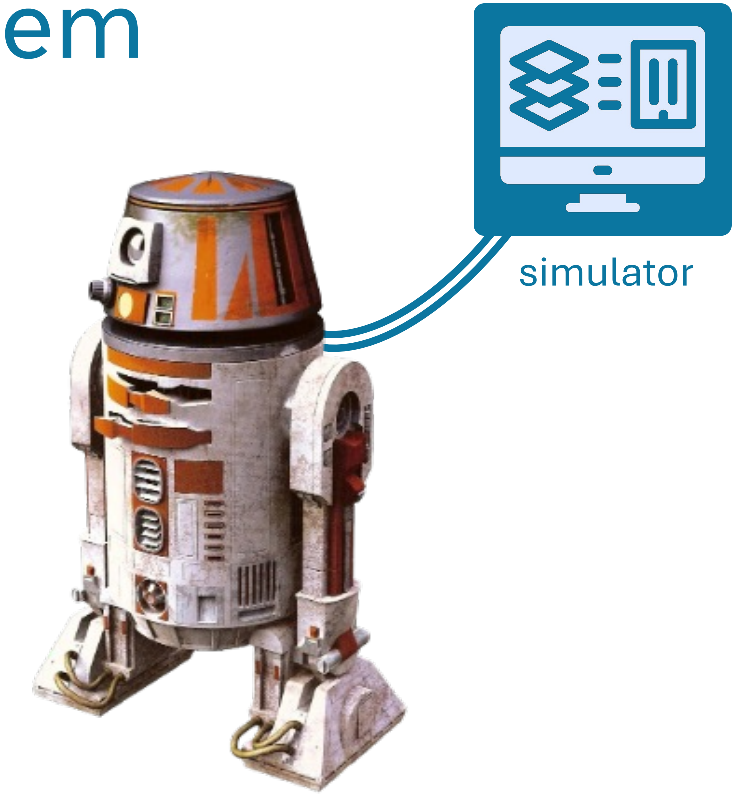
- Can be connected to a simulator.
- Can have arbitrary internal implementation.
- Does not know input vocabulary.

Input

- Predicates (User vocabulary)
  - With their evaluation functions

Output

- A description of agent's working:
  - A list of capabilities.
  - An interpretable description of each capability.



Black-Box AI

Interpretable Description: PDDL/PPDDL

```
(: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

01

Introduction

02

Foundational Approach

03

Generalizations

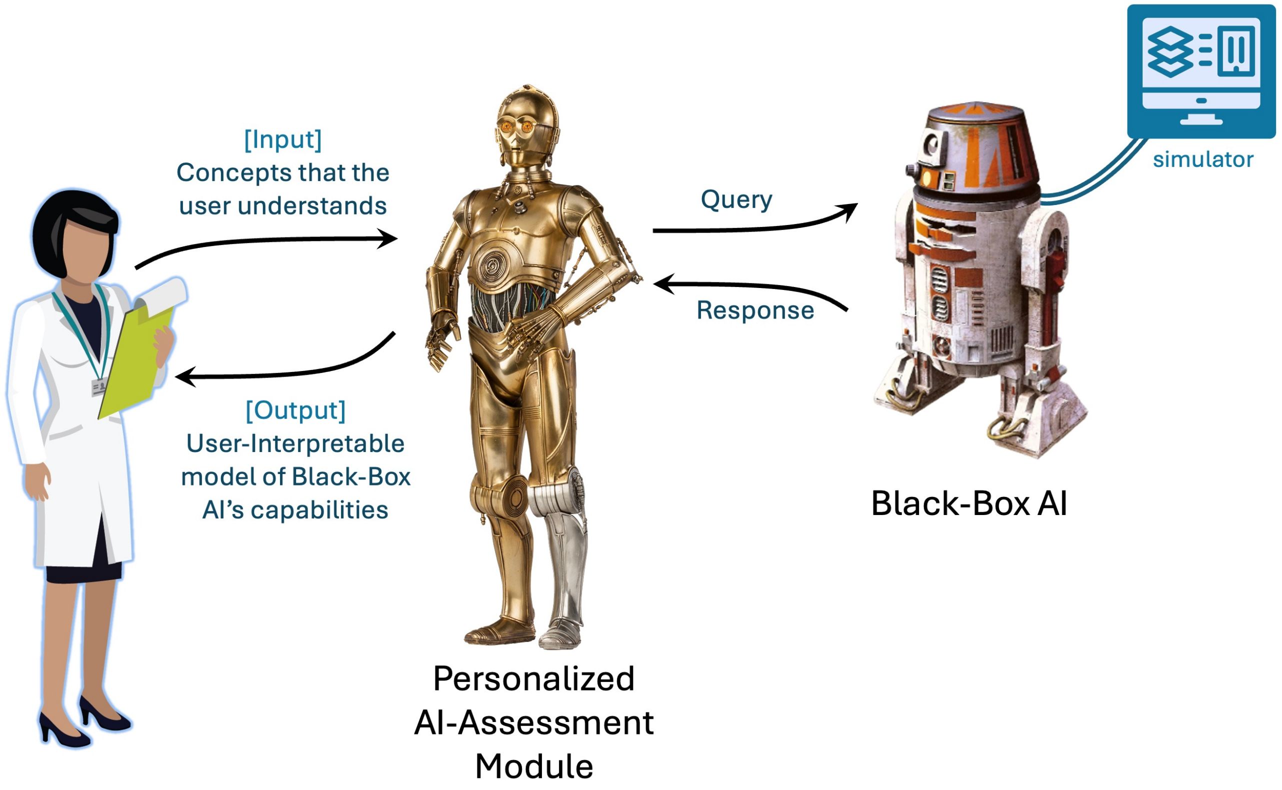
04

Applications

05

Conclusion

Deterministic and Stationary Setting



Assumptions

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

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)`

- $9^{|C|\times|P|}$ = $9^{1\times4}$=6561 possible models (Assuming deterministic models/ descriptions, i.e., no probabilities).

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

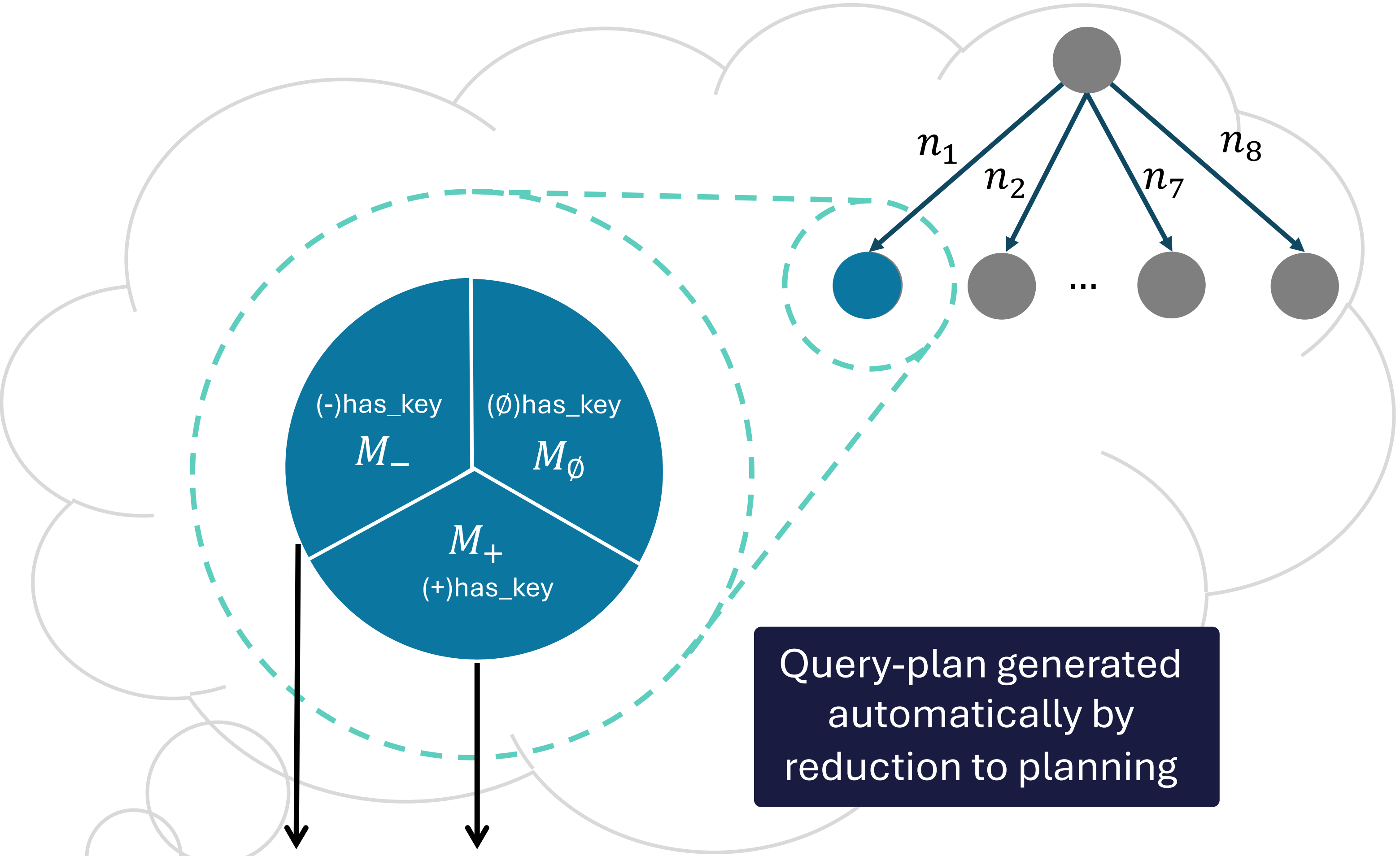
Simple Queries

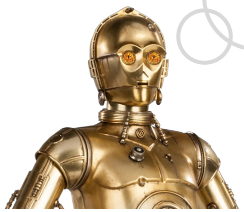
	Plan Outcome Queries	State Reachability Query
Query	In state s_I, what will happen if you execute the plan $\pi = \langle c_1, \ldots, c_n \rangle$?	Can you go from state s_I to state s_F?
Response	I can execute first ℓ steps of the plan, ending up in state s_F.	Yes / No.

- 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



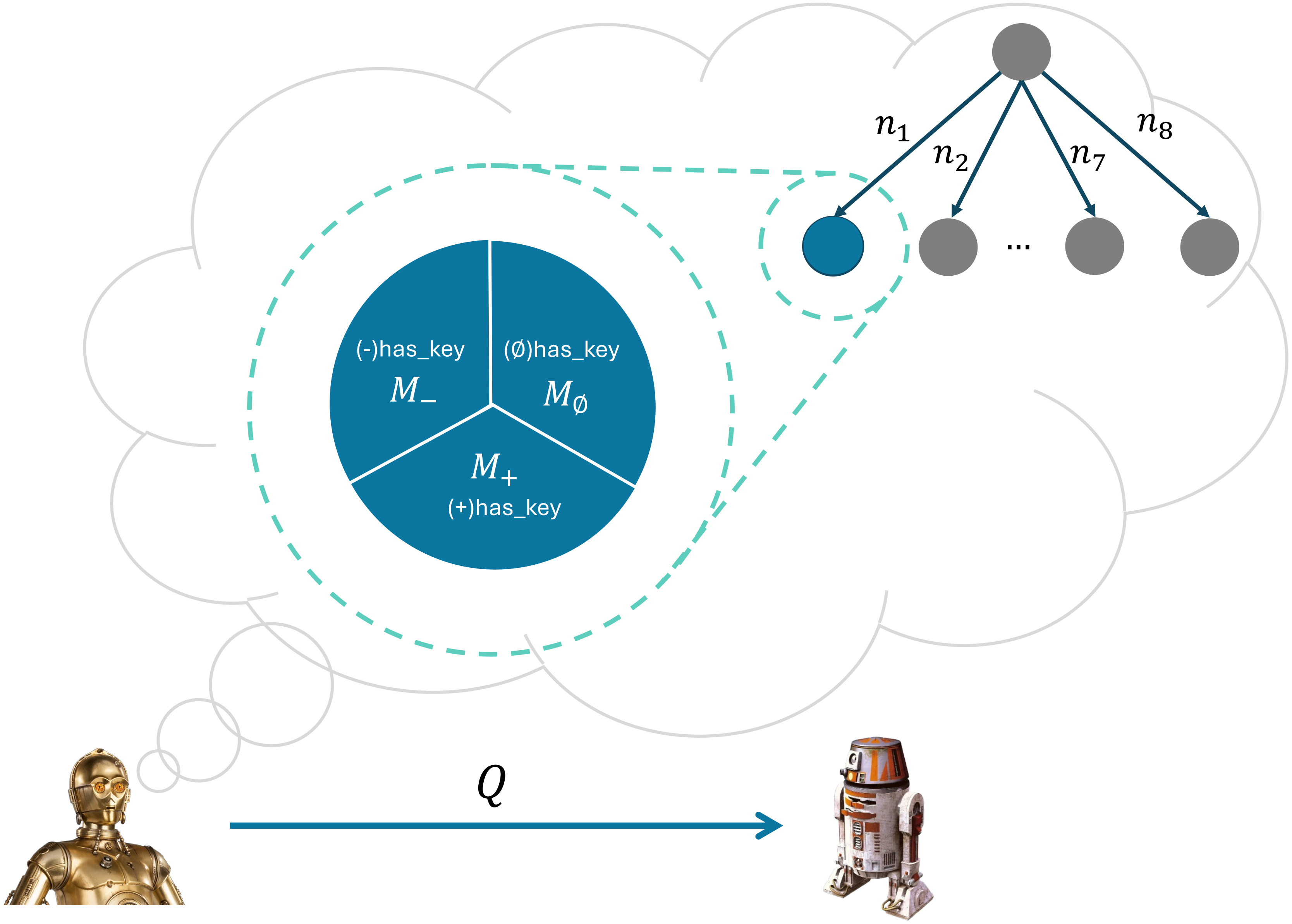


Generate a
distinguishing query:
Q such that $Q(M_-) \neq Q(M_+)$

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

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

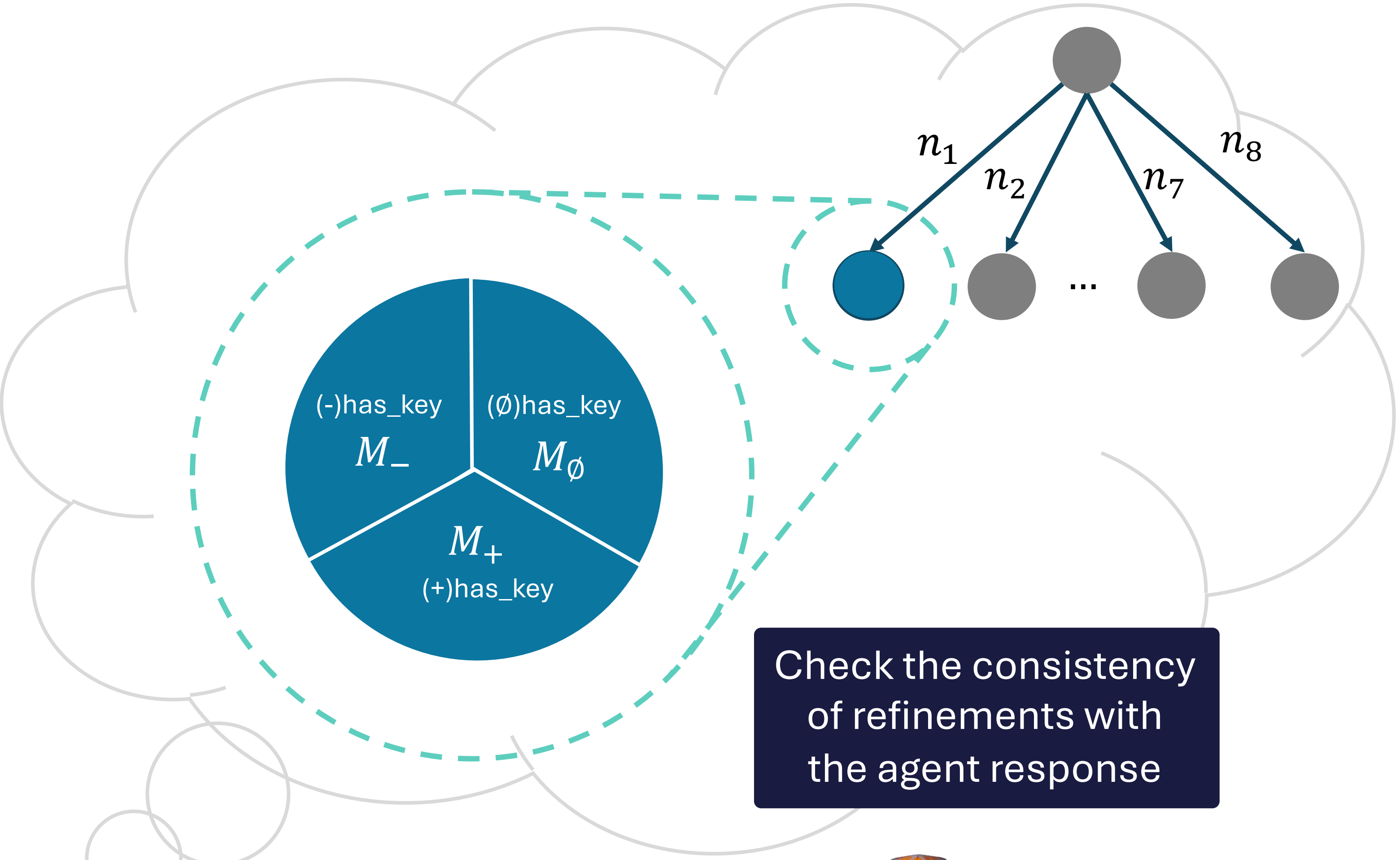
Hierarchical Query Synthesis



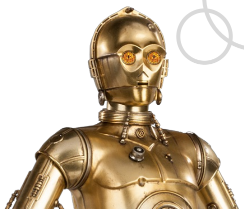

```
(:action open-door
  :parameters (?l1)
  :precondition (and
n1  (+/-/Ø)(has_key)
n2  (+/-/Ø)(door_open)
n3  (+/-/Ø)(door_adjacent ?l1)
n4  (+/-/Ø)(player_at ?l1))
  :effect (and
n5  (+/-/Ø)(has_key)
n6  (+/-/Ø)(door_open)
n7  (+/-/Ø)(door_adjacent ?l1)
n8  (+/-/Ø)(player_at ?l1))
```

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

Hierarchical Query Synthesis



Check the consistency of refinements with the agent response



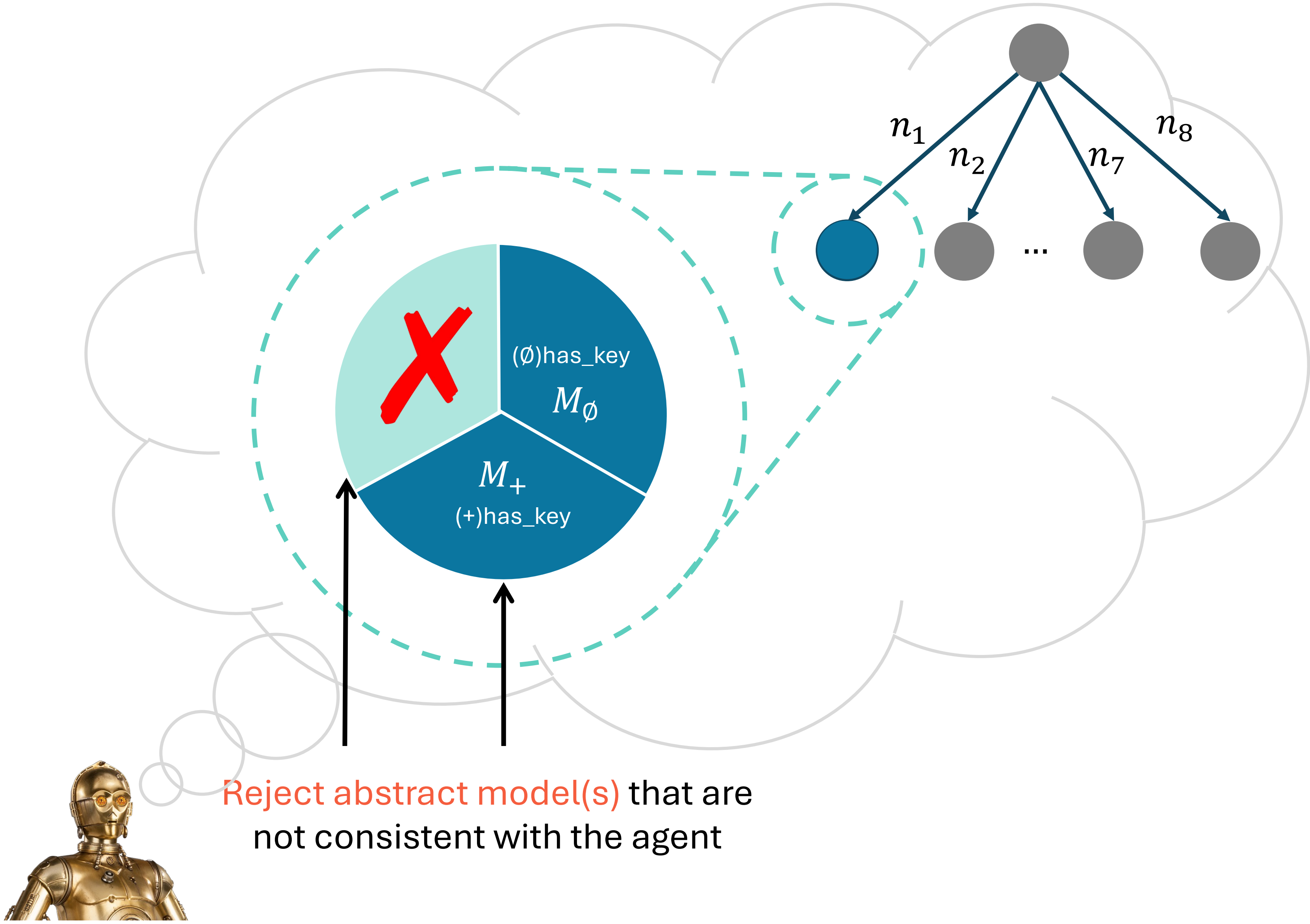
$$\theta = Q(Agent)$$

$$Q(M_-) \neq Q(M_+)$$

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

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

Hierarchical Query Synthesis

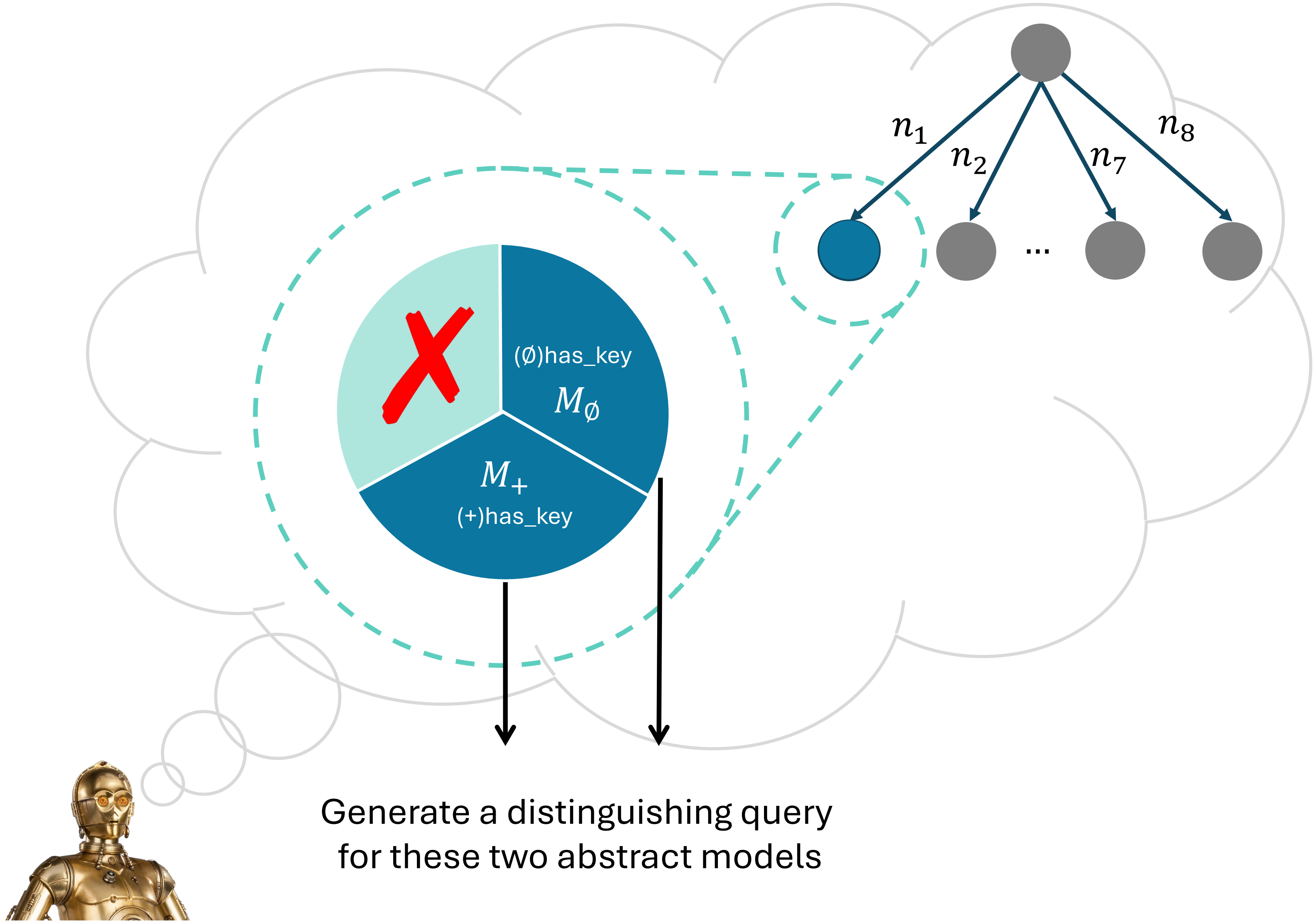


Reject abstract model(s) that are not consistent with the agent

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

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

Hierarchical Query Synthesis

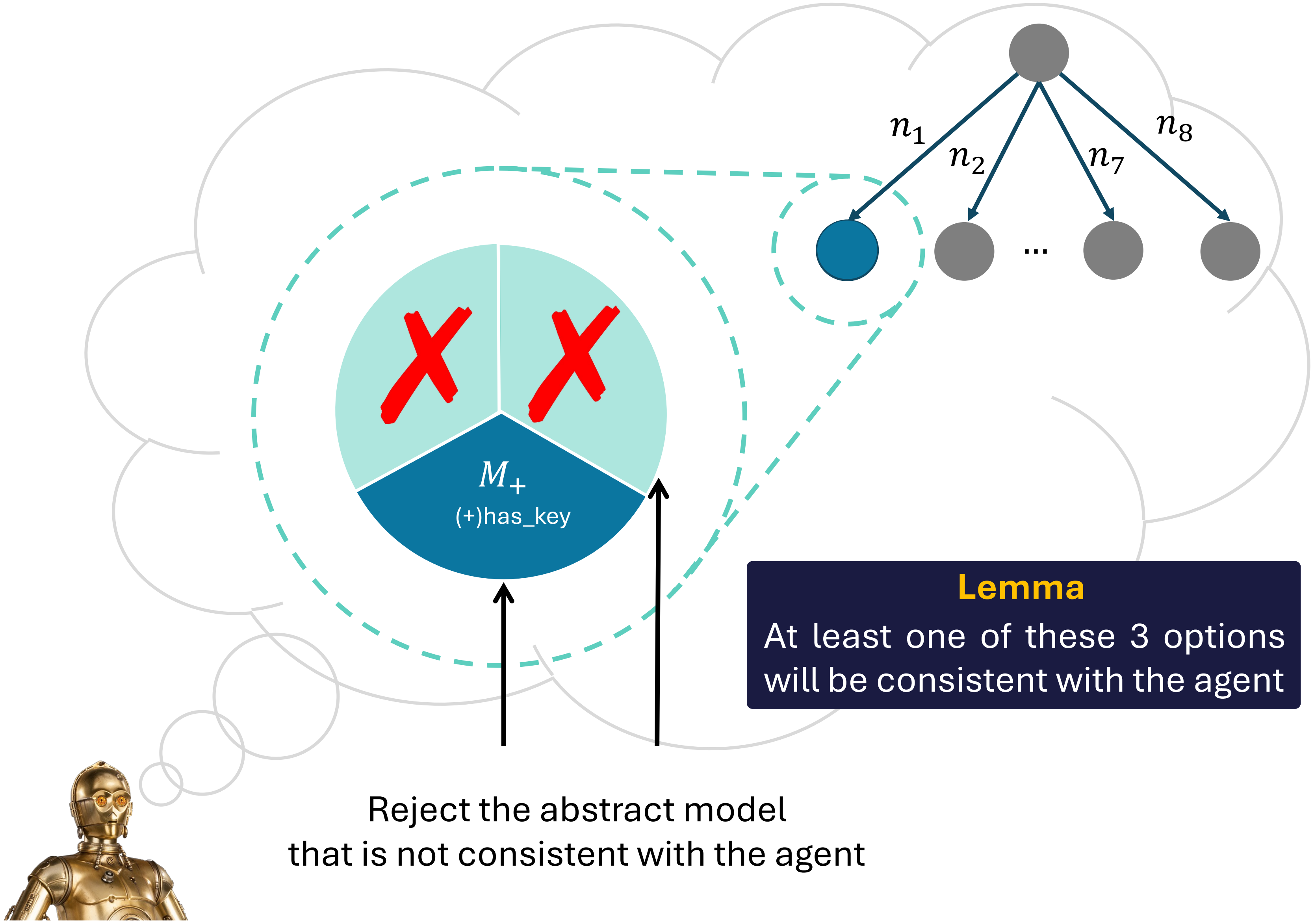


Generate a distinguishing query
for these two abstract models

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

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

Hierarchical Query Synthesis



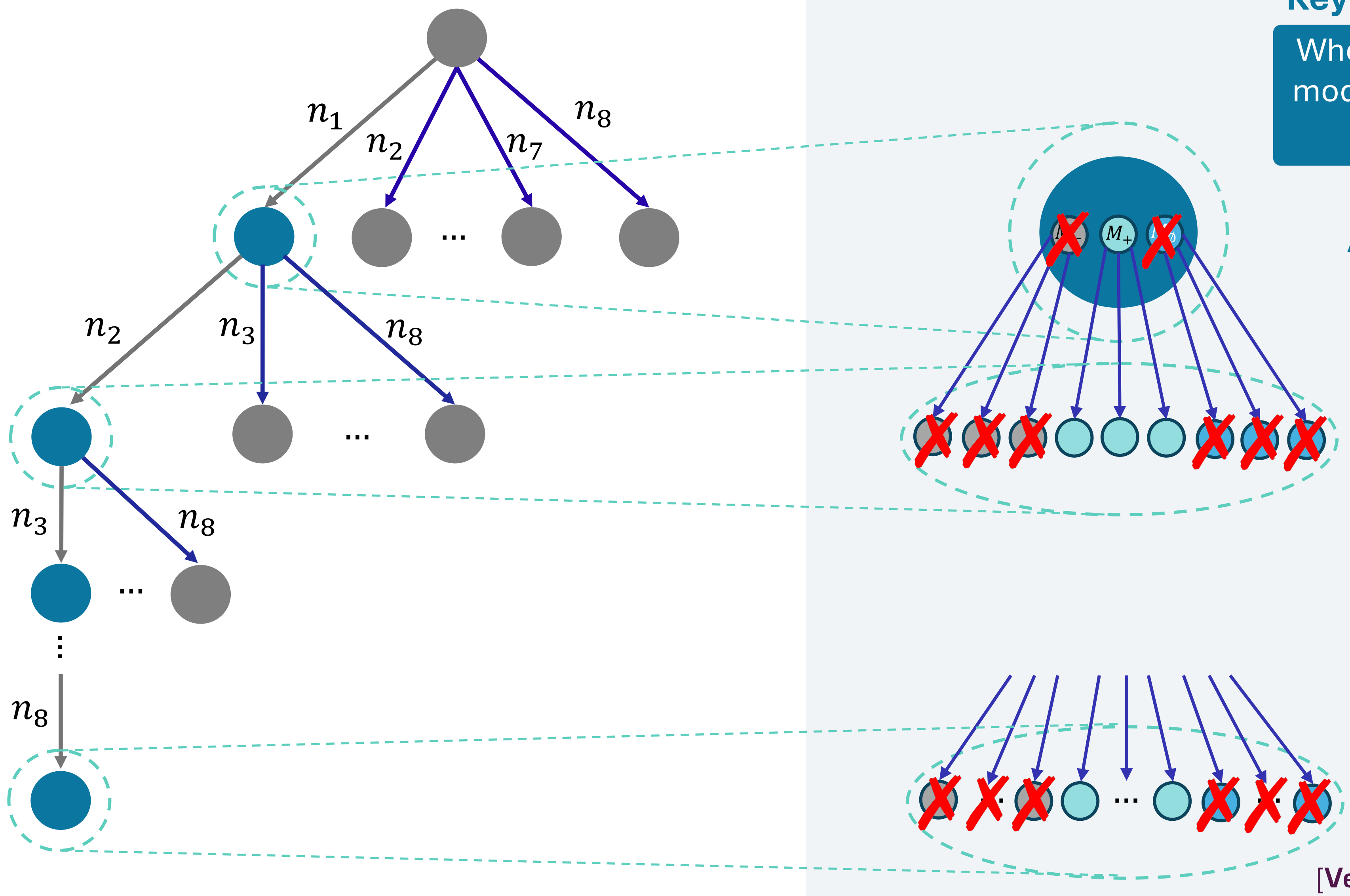
Lemma
At least one of these 3 options will be consistent with the agent

Reject the abstract model
that is not consistent with the agent

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

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

Hierarchical Query Synthesis



Key feature of the algorithm

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

Active Learning

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

Deterministic and Stationary Setting

Input

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

Output

- PDDL-like description of each capability.

Assumptions

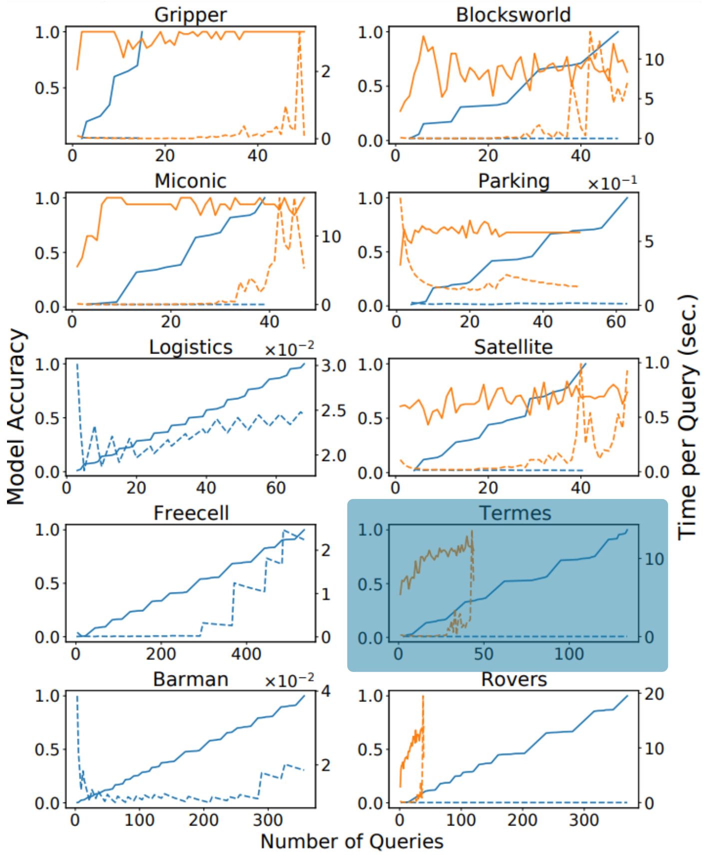
- User's vocabulary matches simulator's vocabulary.
- Black-Box AI provides a list of capabilities.
- Stationary agent model.
- Deterministic environment.
- Fully observable setting.

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

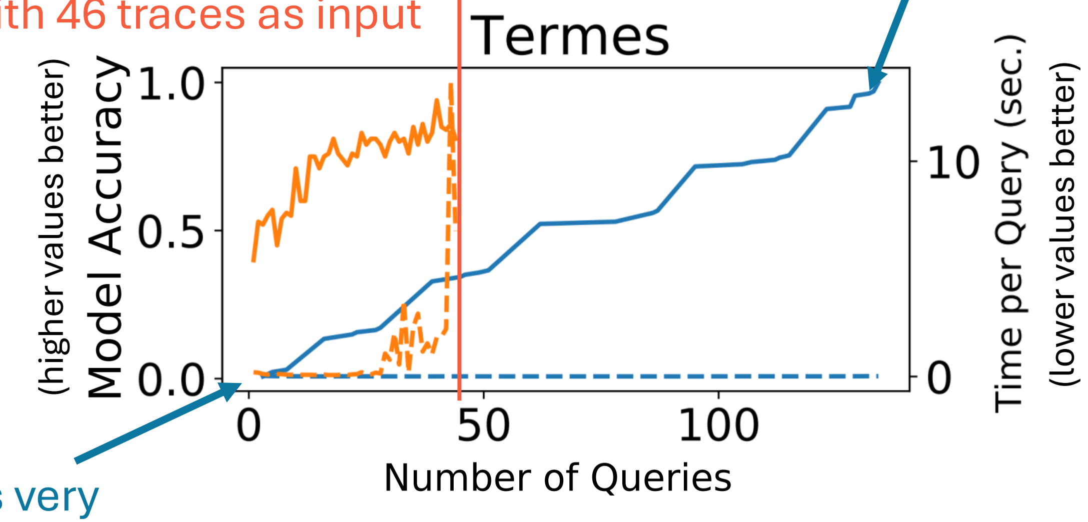
Random deterministic planning agent from IPC

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]

†Aineto, D.; Celorrio, S. J.; and Onaindia, E. 2019. *Learning Action Models With Minimal Observability*. Artificial Intelligence 275: 104–137.

AAM learns Accurate Deterministic Models

- Theorem (*termination*) : The algorithm terminates after a finite number of iterations.

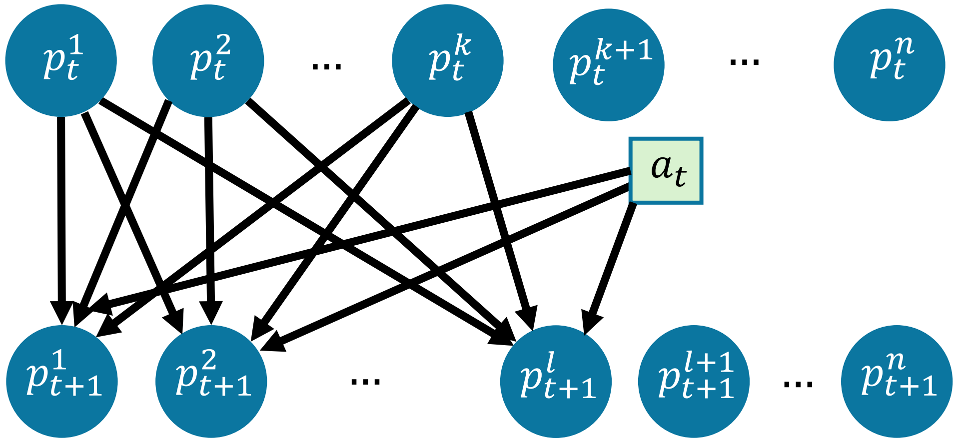
- Theorem (*soundness*): The resulting (set of) model(s) is(are) functionally equivalent to the ground truth model.

Lemma 8 in Thesis

Theorem 4 in Thesis

Causal Accuracy Analysis

- Use the framework for *Actual Causality*† to define the causal accuracy of the models that we learn.
- Explain theoretically why models learned using passive learners may not be causally accurate.
- Show that the models AAM learns are causally accurate†.
  (Theorem 11 in Thesis)



[**Verma**, Srivastava; 2024]

†According to AC2 definition of Actual Cause. Joseph Halpern. *Actual Causality*. MIT Press. 2016.

Stochastic and Stationary Setting

Input

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

Output

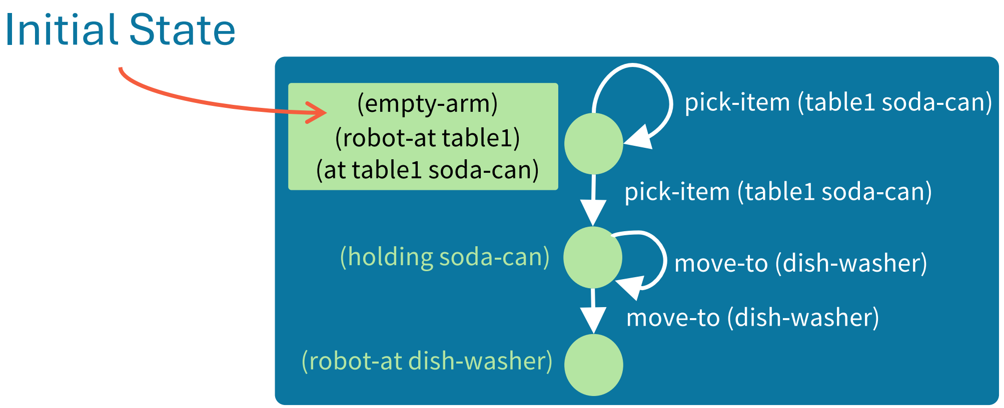
- PPDDL-like description of each capability.

Assumptions

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

Changes for Stochastic Settings

New Queries



Policy: Generated Autonomously by Reduction to Non-Deterministic Planning

What happens if you start in the given initial state and follow this partial policy?

Assumptions

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

Changes for Stochastic Settings

Step 1: Learn a Non-Deterministic Model

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

Apply Maximum Likelihood Estimation

on the observed data (query responses)

Step 2: Convert to Probabilistic Model

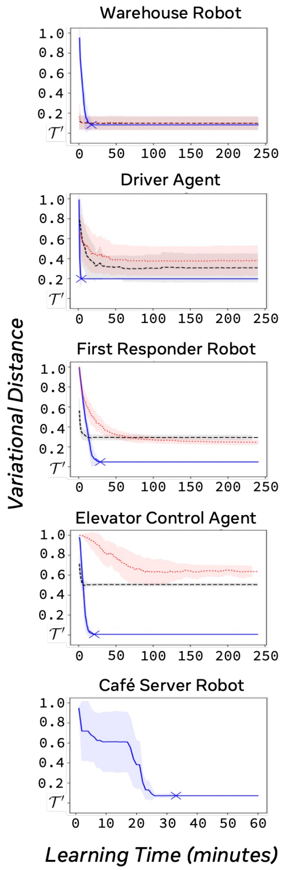
```
(:action open-door
  :parameters (?l1)
  :precondition (and
    (+/-/Ø)(has_key)
    (+/-/Ø)(door_open)
    (+/-/Ø)(door_adjacent ?l1)
    (+/-/Ø)(player_at ?l1))
  :effect (probabilistic
    0.xx (and
      (+/-/Ø)(has_key)
      (+/-/Ø)(door_open)
      (+/-/Ø)(door_adjacent ?l1)
      (+/-/Ø)(player_at ?l1))
    0.yy (and
      (+/-/Ø)(has_key)
      (+/-/Ø)(door_open)
      (+/-/Ø)(door_adjacent ?l1)
      (+/-/Ø)(player_at ?l1)))
```

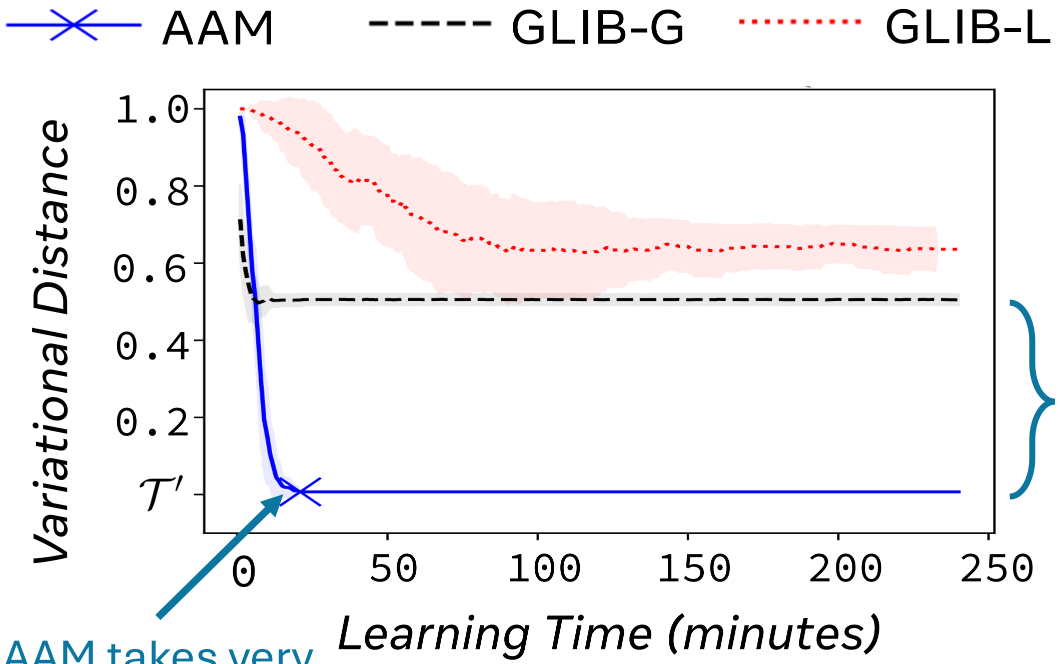
[**Verma**, Karia, Srivastava; NeurIPS '23]

AAM learns accurate probabilistic models faster

- Baseline: directed exploration approach (GLIB)
- Increase time taken to learn the model.

Random probabilistic planning agent from IPC





AAM learns a much better model than GLIB

AAM takes very less time

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

†Chitnis, R.; Silver, T.; Tenenbaum, J.; Kaelbling, L. P.; Lozano-Perez, T. GLIB: Efficient Exploration for Relational MBRL via Goal-Literal Babbling. AAAI 2021.

AAM learns Accurate Probabilistic Models

- Theorem (*soundness and completeness*): The intermediate non-deterministic model (after step 1) is sound and complete w.r.t. the ground truth model.

Theorem 9 in Thesis

- Theorem (*probabilistic correctness*): The resulting probabilistic model is correct w.r.t. the ground truth model.

Theorem 10 in Thesis

01

Introduction

02

Foundational Approach

03

Generalizations

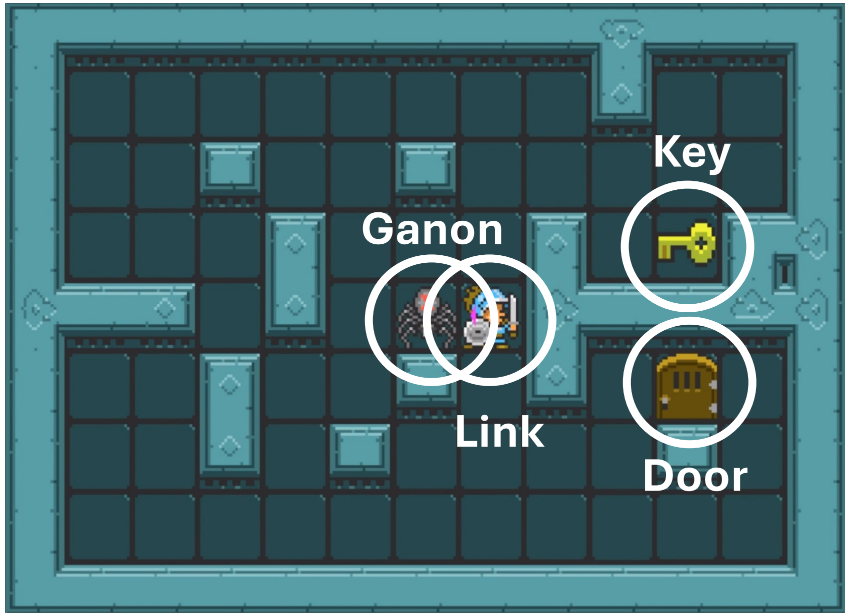
04

Applications

05

Conclusion

User Vocabulary can be Less Expressive



Agent's State Representation	State Representation in User's Vocabulary
pixel_1_1(#42A8B3)	(at ganon 5,3)
pixel_1_2(#42A8B3)	(at link 6,3)
.	(at key 9,4)
.	(at door 9,2)
.	
pixel_n_m(#203A3D)	

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

Discovering Capabilities

Input

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

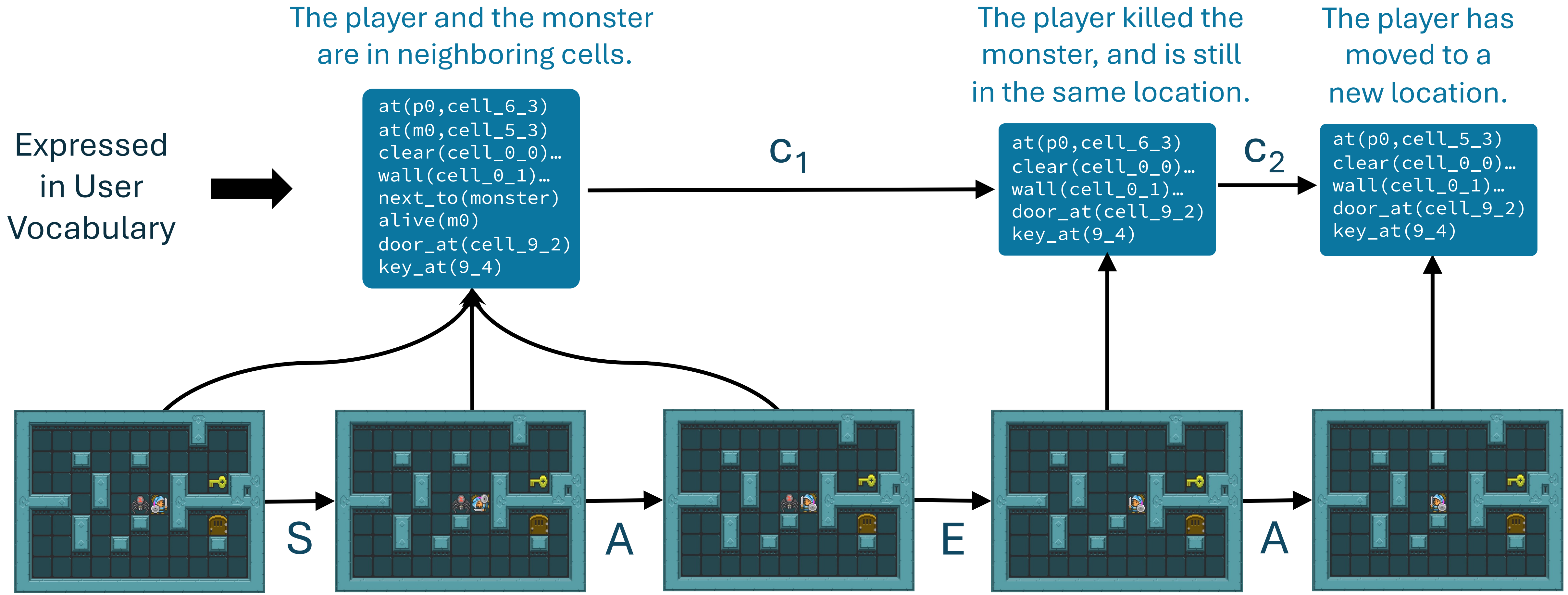
Output

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

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.

Discovering Capabilities using Input Predicates as Abstractions



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

Example of a Learned Capability Description

```
(: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

This capability is: "Defeat Ganon"

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

User Study Setup to Verify Interpretability

Preconditions

4. **Capability C4**:
The *player* can execute this capability when:

- The *monster* is not defeated.
- The *player* is in *cell1*.
- The monster is in *cell2*.
- The player is in a cell adjacent to the *monster*.

Effects

After the *player* executes this capability:

- *Cell2* is empty.
- The *monster* is defeated.
- The *monster* is not in *cell2*.
- The player is not in a cell adjacent to the *monster*.

Question 4 of 12:
Select the phrase that best summarizes the capability **C4**? We will use your response while referring to this capability **C4** later in the survey.

Go next to Door
Go next to Ganon
Go next to Key
Go next to Wall
Defeat Ganon
Break Key
Pick Key
Open Door

Possible options to choose from

[Capability Description Example]

Keystroke Description

W: Pressing this key does the following:

- If Link is facing up and there is no wall, door, or key in the cell above, then Link moves to the cell above.
- If there is a wall, door, or key in the cell above Link, then Link stays in the same cell.
- If Link is facing Left, Right, or Down before pressing W, then Link faces up but stays in the same cell.

Question 1 of 11:
Select the phrase that best summarizes pressing **W**? We will use your response while referring to this key **W** later in the survey.

Up
Down
Left
Right
Interact

Possible options to choose from

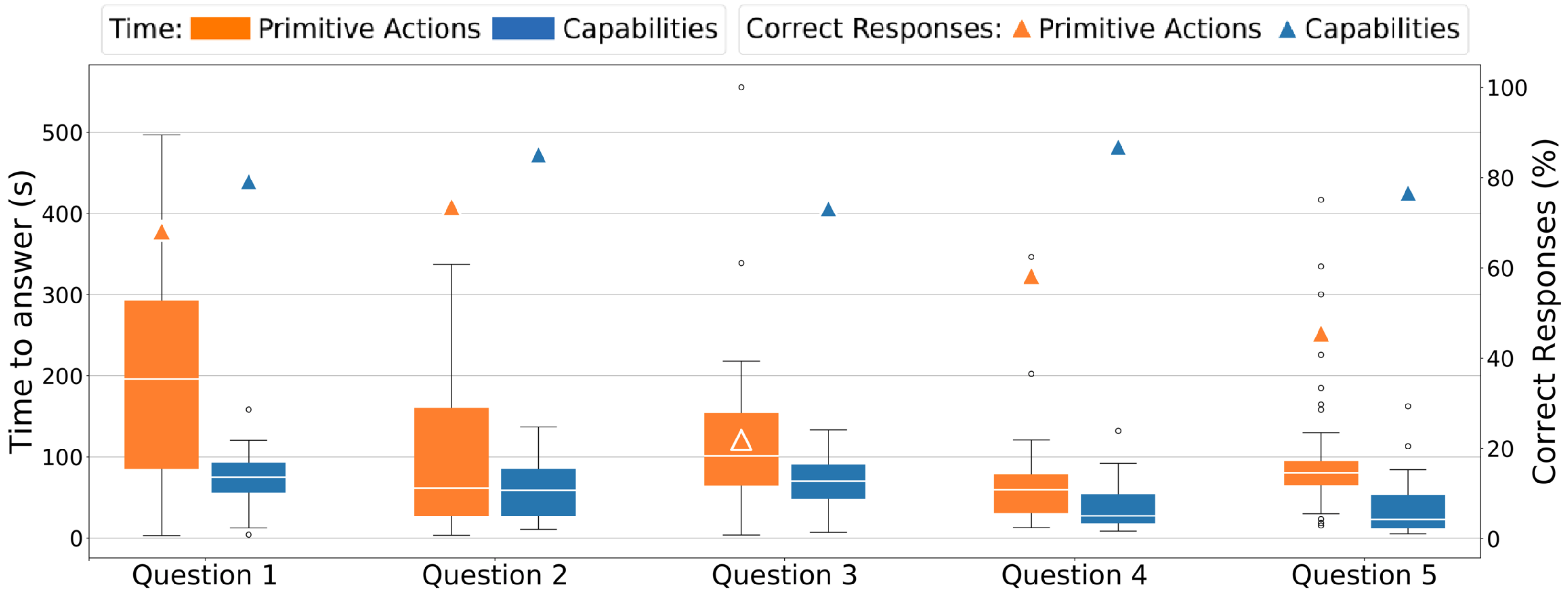
[Functionality Description Example]

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

Utility of Discovered Capability Descriptions

If Link starts in the state shown below:

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



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

Learned Capability Descriptions are Maximally Consistent

- Theorem (*consistency*): The learned descriptions are consistent with the observations and the queries.

Theorem 5 in Thesis

- Theorem (*maximal consistency*): 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.

Theorem 6 in Thesis

- Theorem (*probabilistic completeness*): In the limit of infinite execution traces, the probability of discovering all capabilities expressible in the user vocabulary is 1.

Theorem 8 in Thesis

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

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.

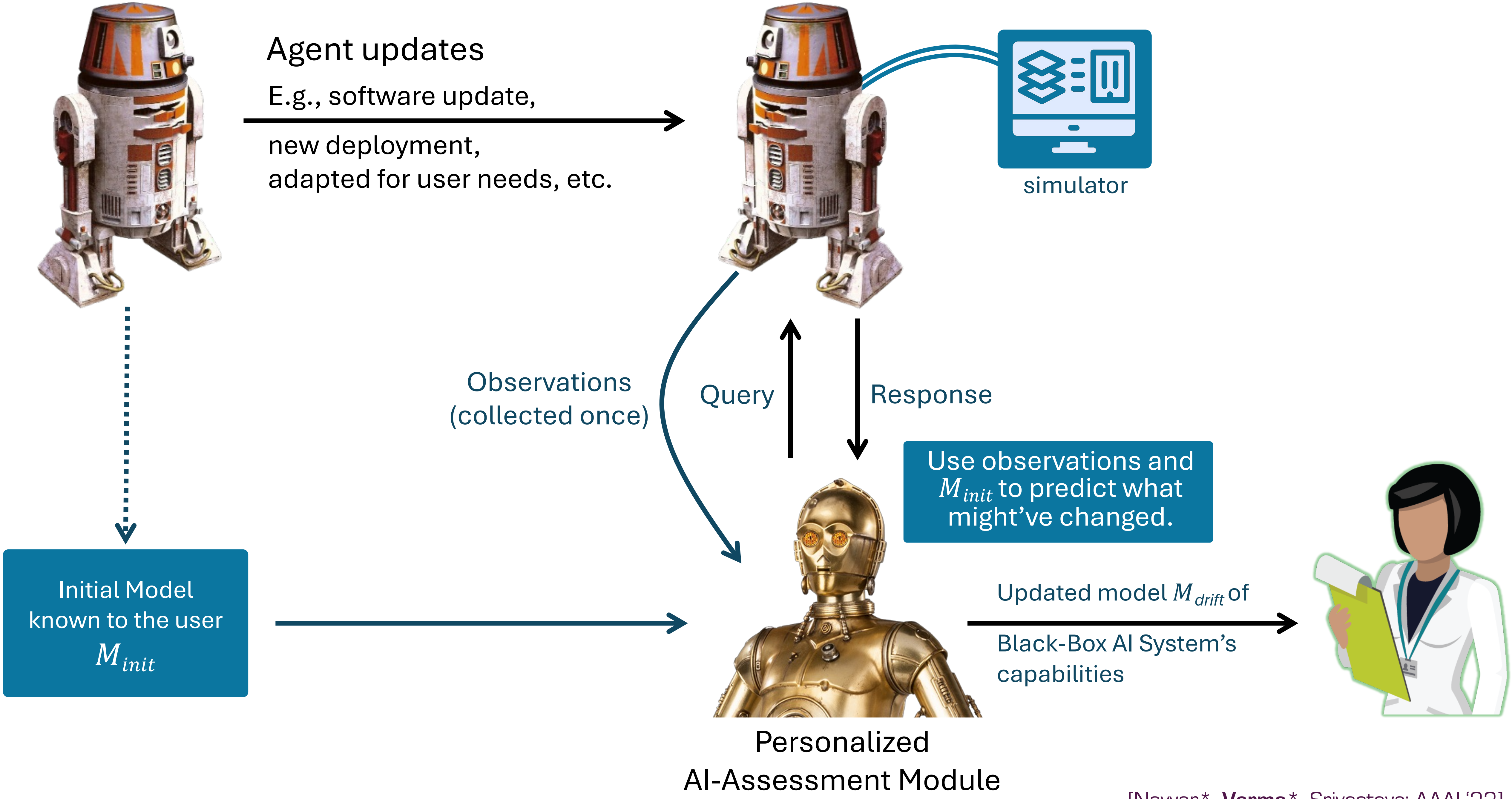
Output

- Updated PDDL-like description of each capability.

Can we learn an updated model without doing a complete assessment?

Assumptions

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

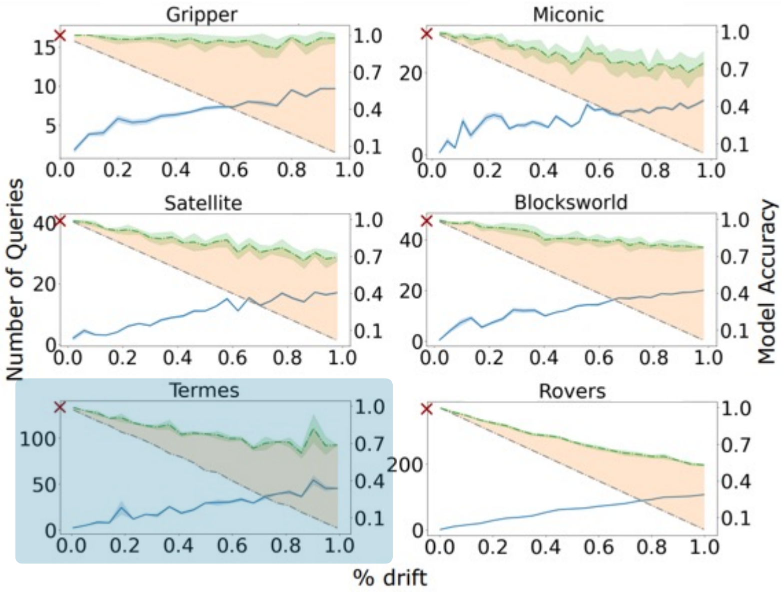


[Nayyar*, **Verma***, Srivastava; AAAI '22]

Fewer Queries Needed Compared to Learning from Scratch

Random deterministic planning agent from IPC

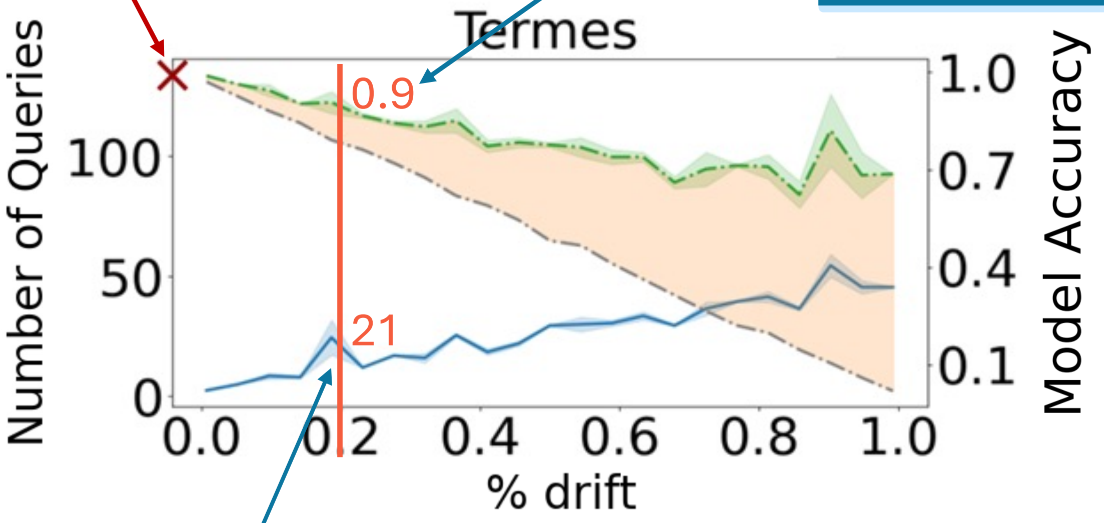
Accuracy of initial model
Accuracy gained by AAM
Accuracy of model computed by AAM
Number of queries by AAM
Number of queries when learning from scratch



134 queries needed if starting from scratch

Accuracy

Used 10 observations per domain



Number of queries are much lower than 134

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

Learned Updated Capability Descriptions are Consistent

- Theorem (*consistency*): The learned descriptions are consistent with the observations and the query responses.

Theorem 4 in Thesis

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

01

Introduction

02

Foundational Approach

03

Generalizations

04

Applications

05

Conclusion

Continual Learning and Planning (CLaP)

- Applying Agent Assessment to RL settings.
  - Agent does not know the model.
  - List of capabilities is known.
  - List of predicates is known.

- Learning a model for both agent and environment.

- Use assessment to see how the environment responds to agent actions.

Setting

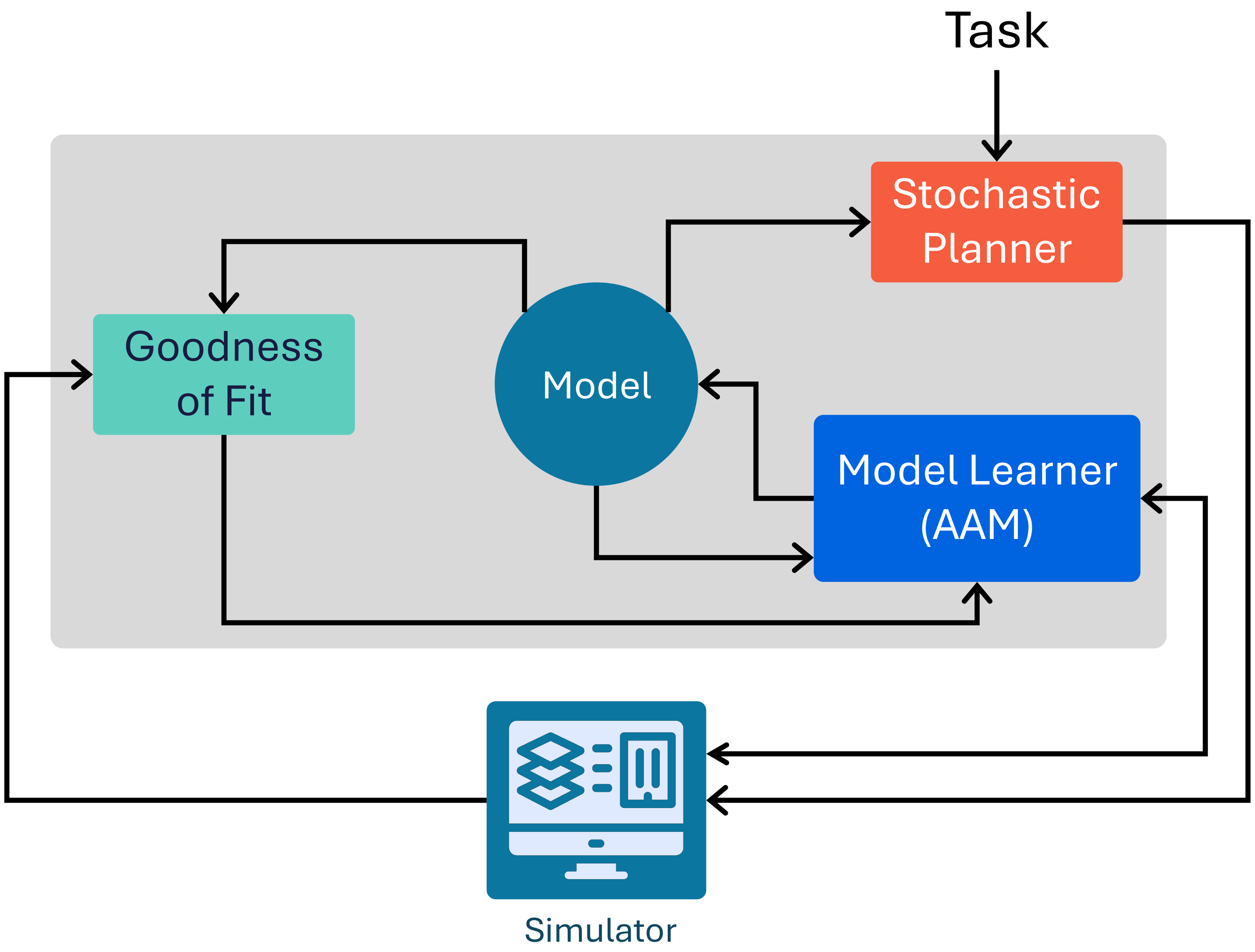
- A stream of tasks as input.
- Different goals for each task.
- Simulator's transition function can change arbitrarily.

Objective

- Maximize #tasks completed within a fixed budget.
- Minimize adaptive delay and regret.

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

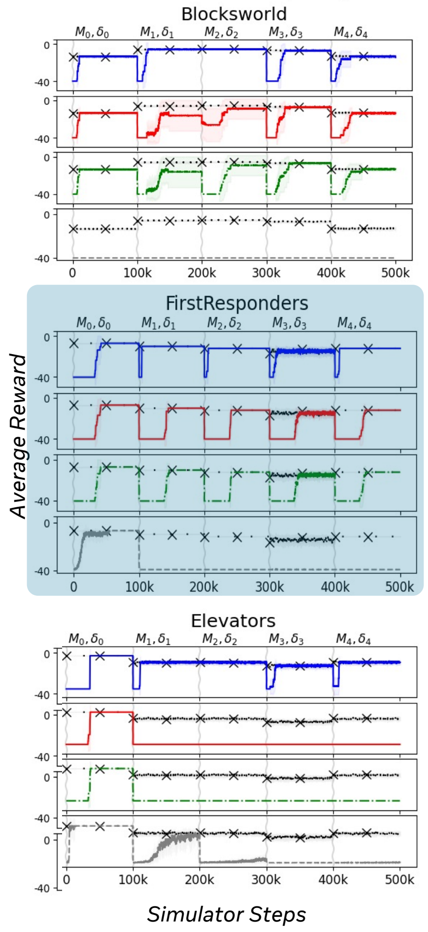
Continual Learning and Planning (CLaP)



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

CLaP Few Shot Transfers in Non-Stationary Settings

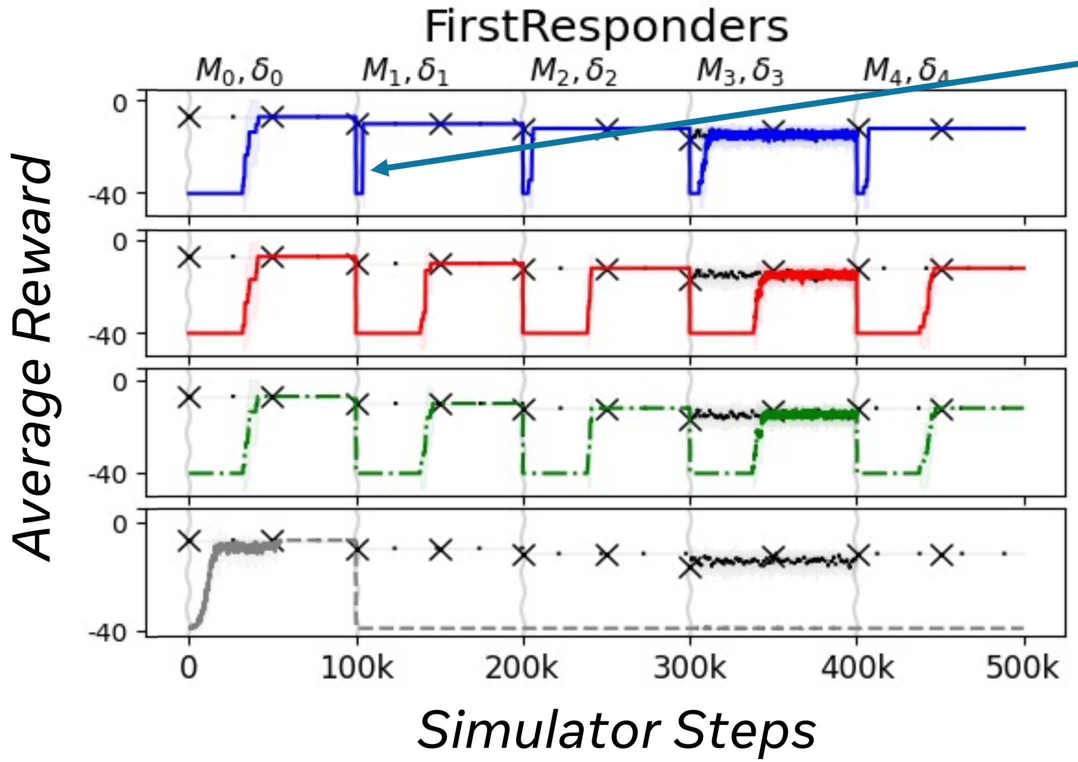
Random probabilistic planning agent from IPC



Oracle
Q-Learning
Adaptive + Need-Based (Ours)
Non-adaptive + Comprehensive
Adaptive + Comprehensive

Ablations

Adaptive Delay least with CLaP



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

01

Introduction

02

Foundational Approach

03

Generalizations

04

Applications

05

Conclusion

Desiderata of Assessment System

Interpretability

Correctness

Generalizability

Easy to Satisfy Requirements

How well we did on Desiderata?

Interpretability	We showed with a user study that the discovered capability models and their descriptions we learn are interpretable.
Correctness	We defined how correctness can be measured for each work and proved that we can achieve it.
Generalizability	The approaches are applicable to any taskable AI that satisfies the given assumptions for each approach.
Easy to satisfy requirements	The requirements on the AI system are: • Simulator access. • Support for simple queries available to any SDM system.

Contributions

- Formally defined the Third-Party AI Assessment Problem for the Taskable AI Systems.
- The first work that shows we can make some assumptions about the interface and assess black-box AI systems on the fly.
- We explain how to assess an adaptive agent after it is deployed.
- Explain theoretically why models learned using passive learners may not be causally accurate.

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; In Preparation]

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).

  Winner of Best Demo Award
- 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.

**Equal Contribution*

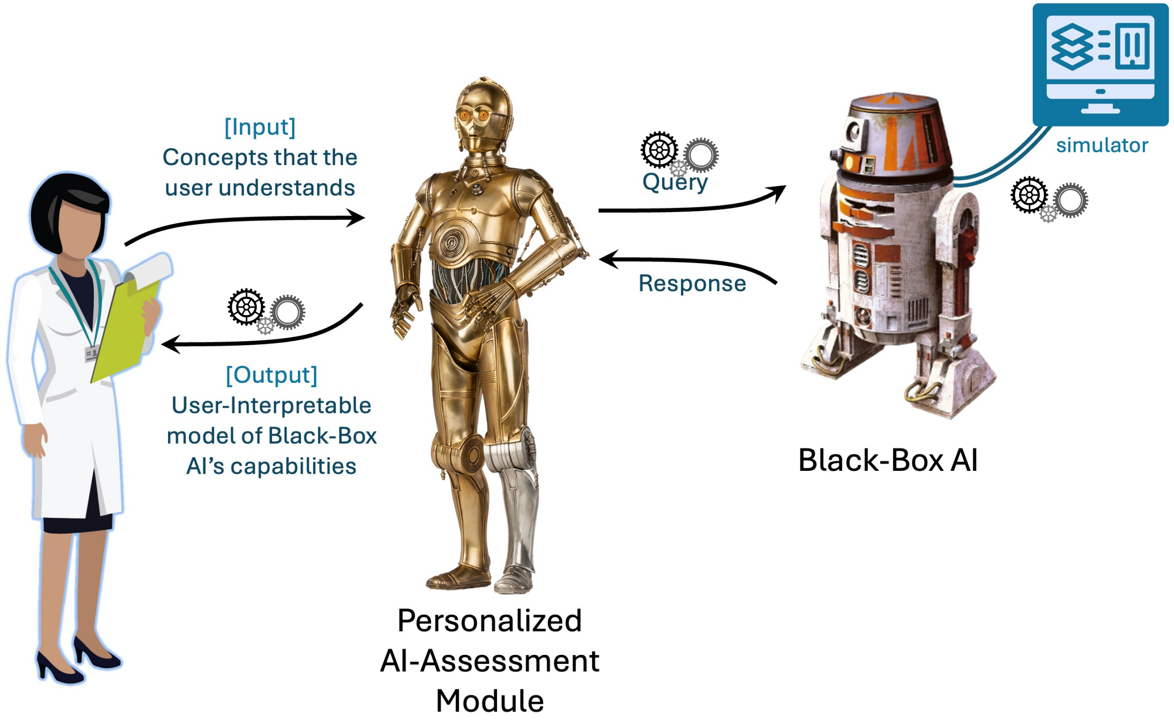
Workshops, Symposia, Preprints

- ***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.

Future Work

- Perform extensive analysis of queries in terms of:
  - Complexity of generating them.
  - Complexity of answering them.
  - Complexity of inferring models from Black-Box AI's responses.

- Extend the work for partially observable settings.




My thesis is written in
ow!
Blood
Sweat
what am i doing with my life??
Tears
JORGE CHAM © 2013
and COFFEE.
WWW.PHDCOMICS.COM

AAIR
Autonomous Agents and Intelligent Robots
Arizona State University

Autonomous Agents and Intelligent Robots Lab


Siddharth Srivastava


Midhun P M


Naman Shah


Kislay Kumar


Chirav Dave


Daniel Molina


Rushang Karia


Rashmeet K Nayyar


Abhyudaya Srinet


Deepak K V


Shashank R Marpally


Mehdi Dadvar


Kiran Prasad


Trevor Angle


Kyle Atkinson


Dylan Fulop


Daniel Bramblett


Gaurav Vipat


Jayesh Nagpal


Daksh Dobhal


Shivanshu Verma


Nancy Cooke


Georgios Fainekos


Yu Zhang


Alberto Speranzon


Subbarao Kambhampati


Sachin Grover


Rohan Chitnis


Alborz Geramifard


Nitin Kamra


Harshit Sikchi


Shentao Yang


J. Benton


Bob Morris


Sydney Wallace


Sarath Sreedharan


Yantian Zha


Swaroop Mishra


Neeraj Varshney


Akku Hanni


Mihir Parmar


Paras Sheth


Mudit Verma


Ramanuj Bhattacharjee


Vinodhini S D


Alberto Olmo


Tejas Gokhale


Kuntal Pal


Adithya Raju


Mirali Purohit


Maitreya Patel


Avisha Das


Anjana Arunkumar


Tasneema Azad


Kevin Vora


Garima Agrawal


Suraj Unni


Aranyak Maity


Shubhodeep Mitra


Karthik Valmeekam


Siddhant Bhambri


Pratanu Mandal

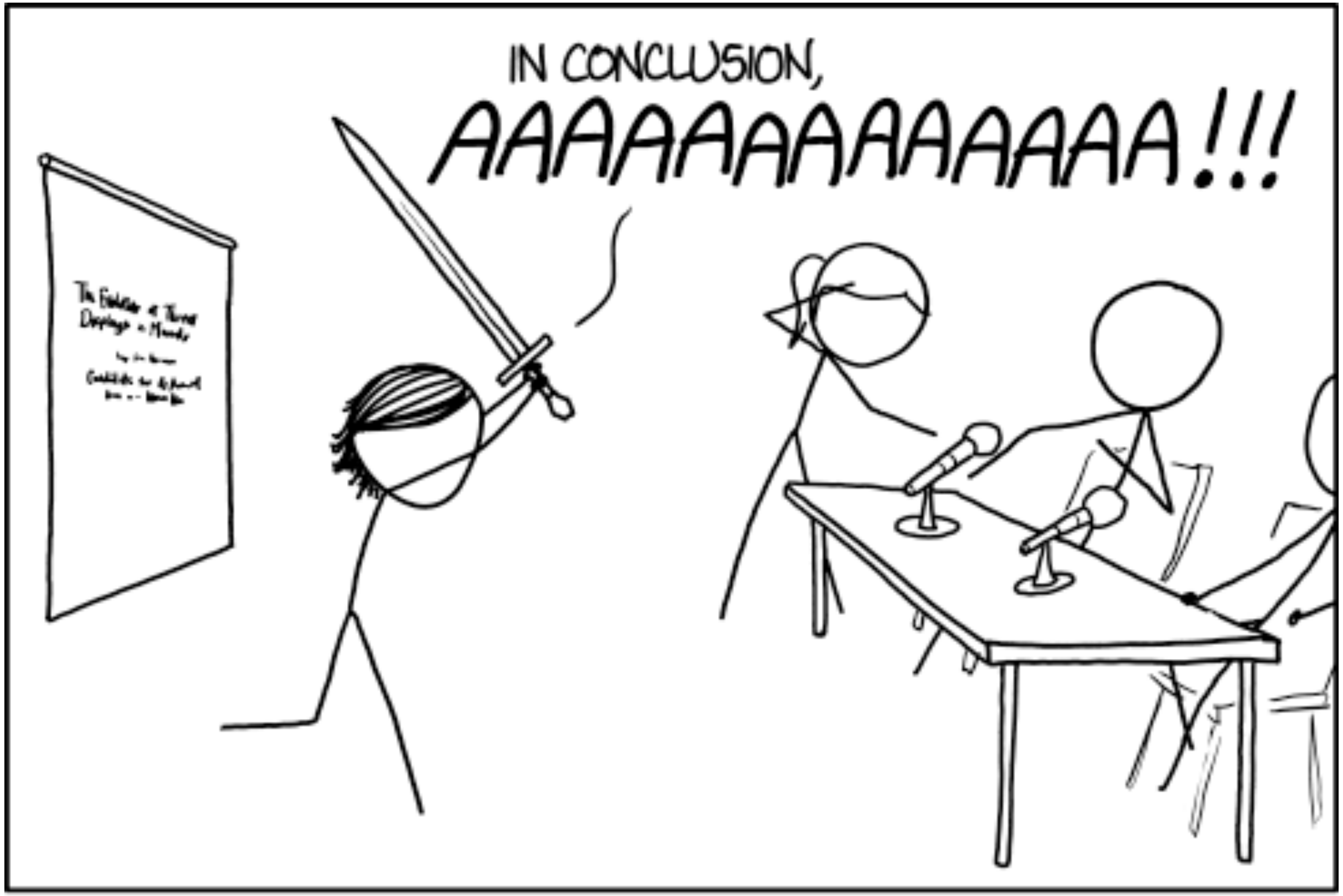

Maitry Trivedi


Angad Kalra


Ahmet Kapkic


Man Luo


Malpooa



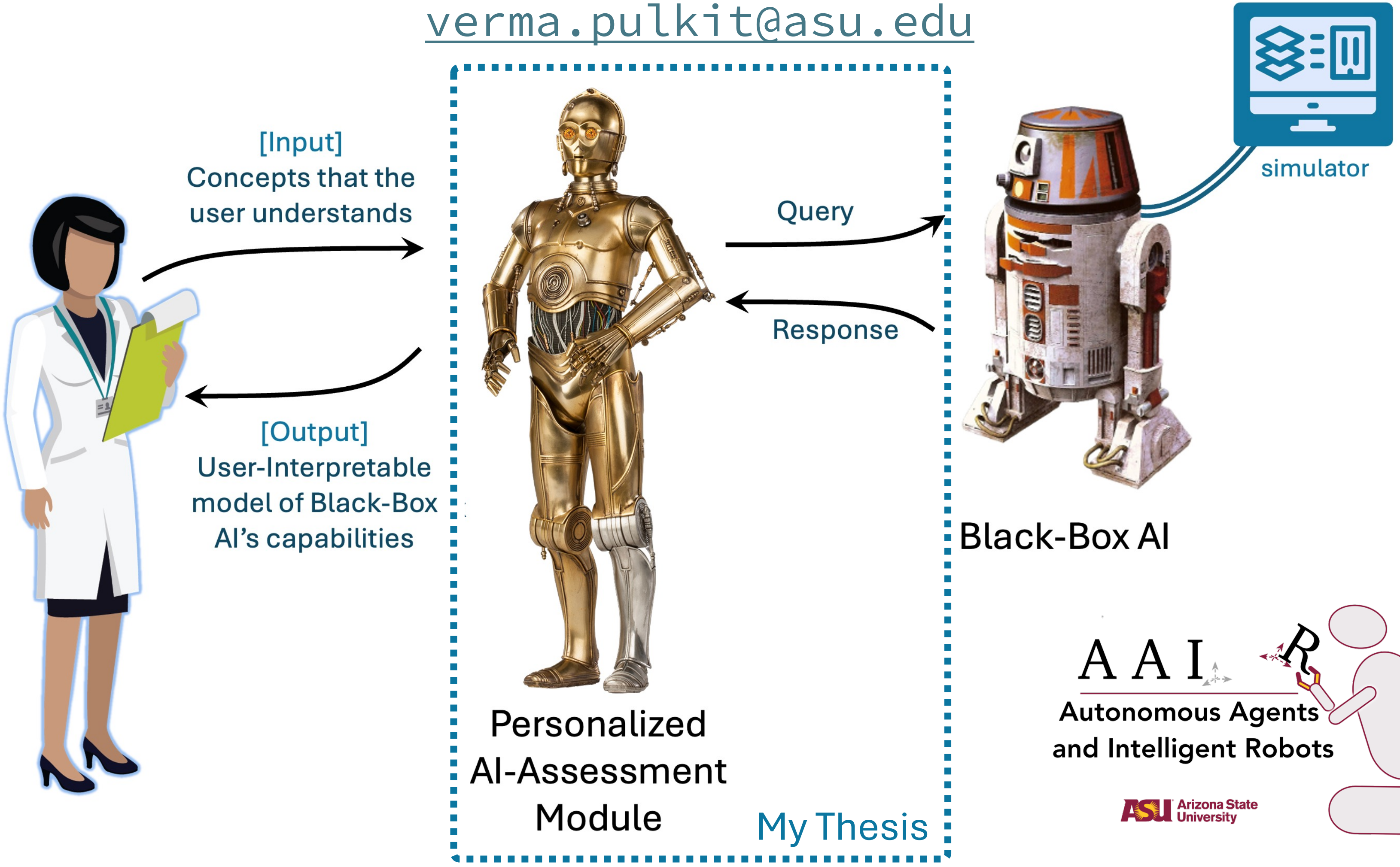
THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

Source: xkcd.com

Data Efficient Paradigms for Personalized Assessment of Taskable AI Systems

Pulkit Verma

verma.pulkit@asu.edu



AAIR
Autonomous Agents and Intelligent Robots

Arizona State University