Talk @ IIT Tirupati – Department of CSE

April 10, 2026

Learning Interpretable Models for Autonomous Assessment of Taskable AI Systems

Pulkit Verma

pulkitv@cse.iitm.ac.in

pulkitverma.net

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.

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

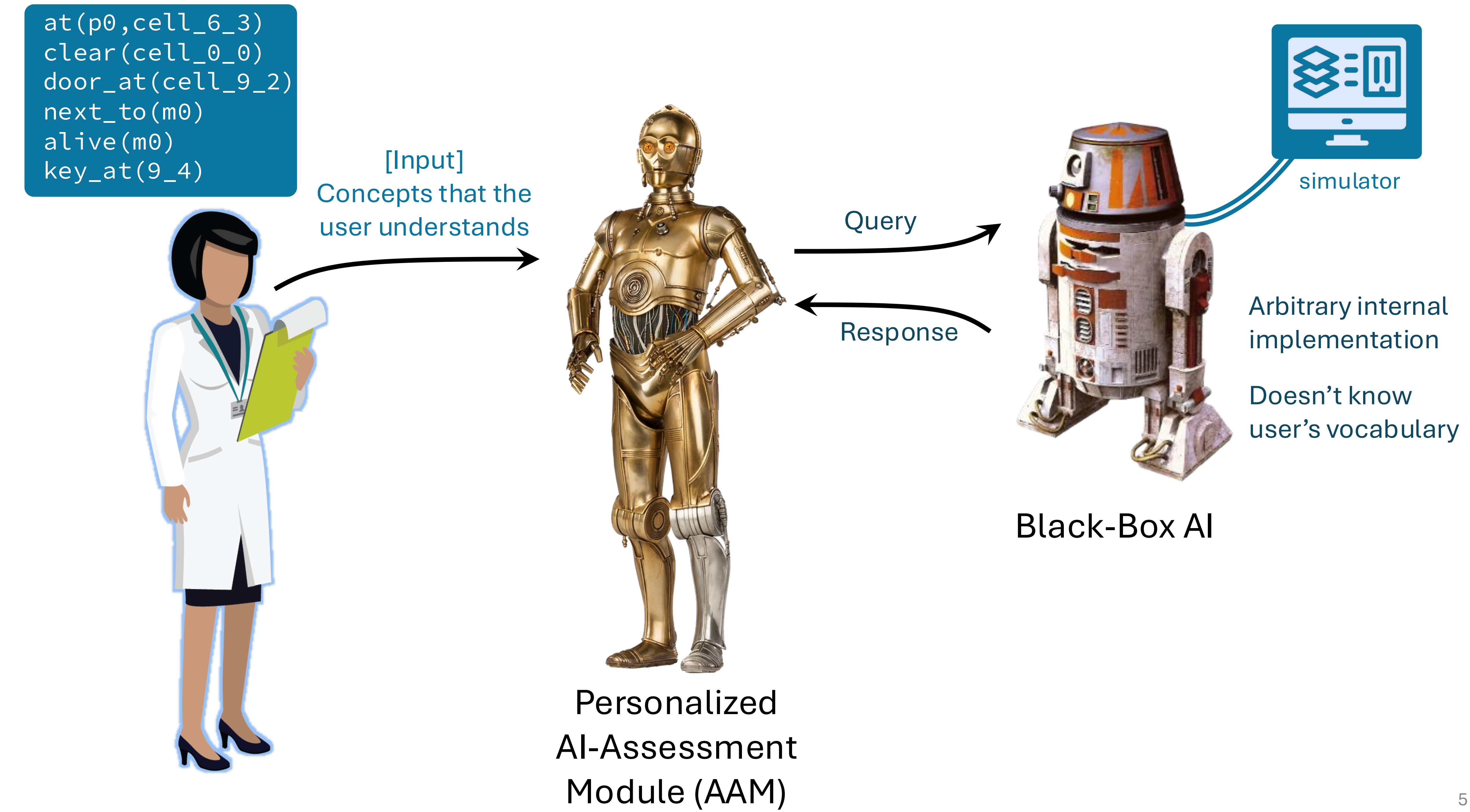
[Input]
Concepts that the
user understands

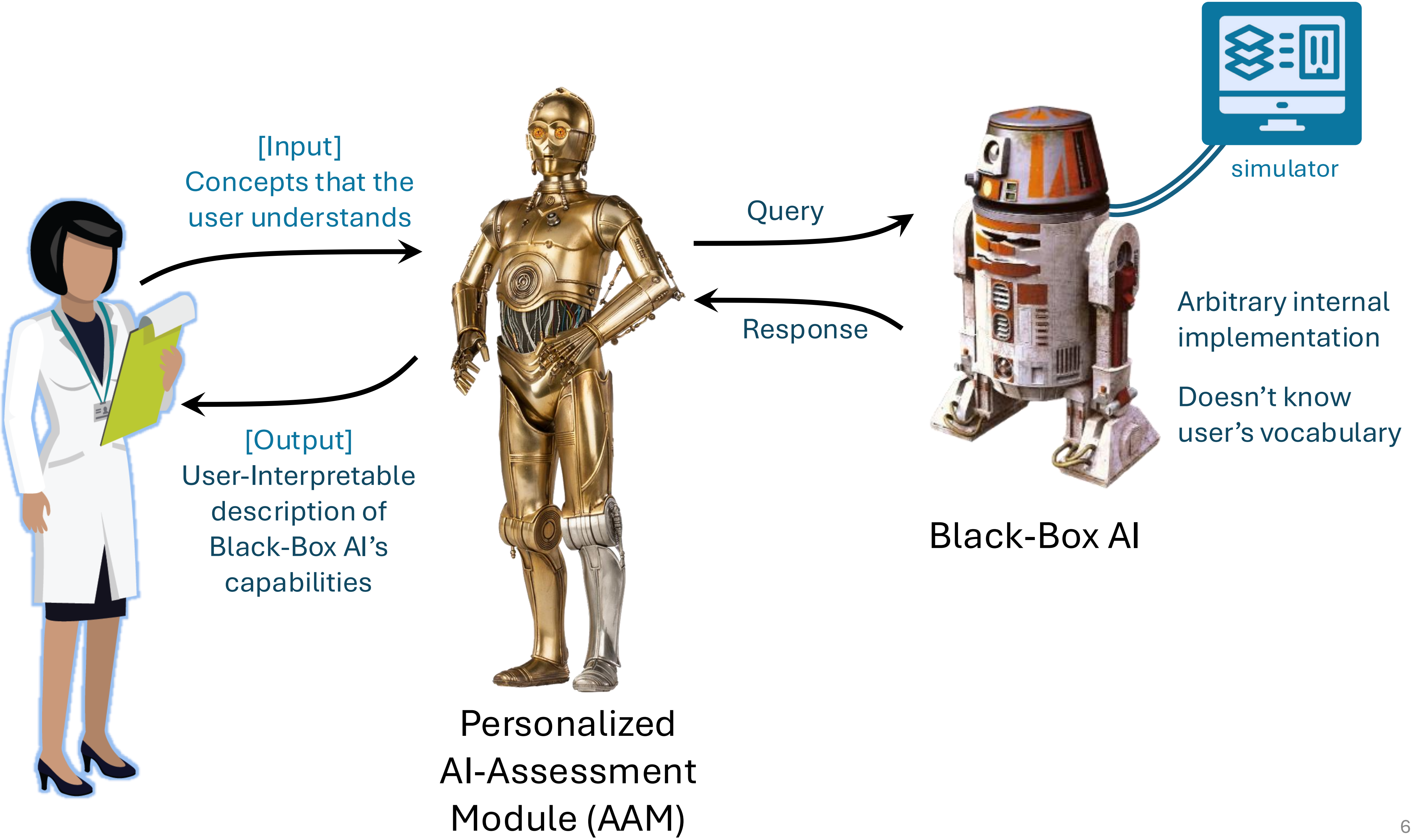
Personalized
AI-Assessment
Module (AAM)

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 (AAM)


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

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

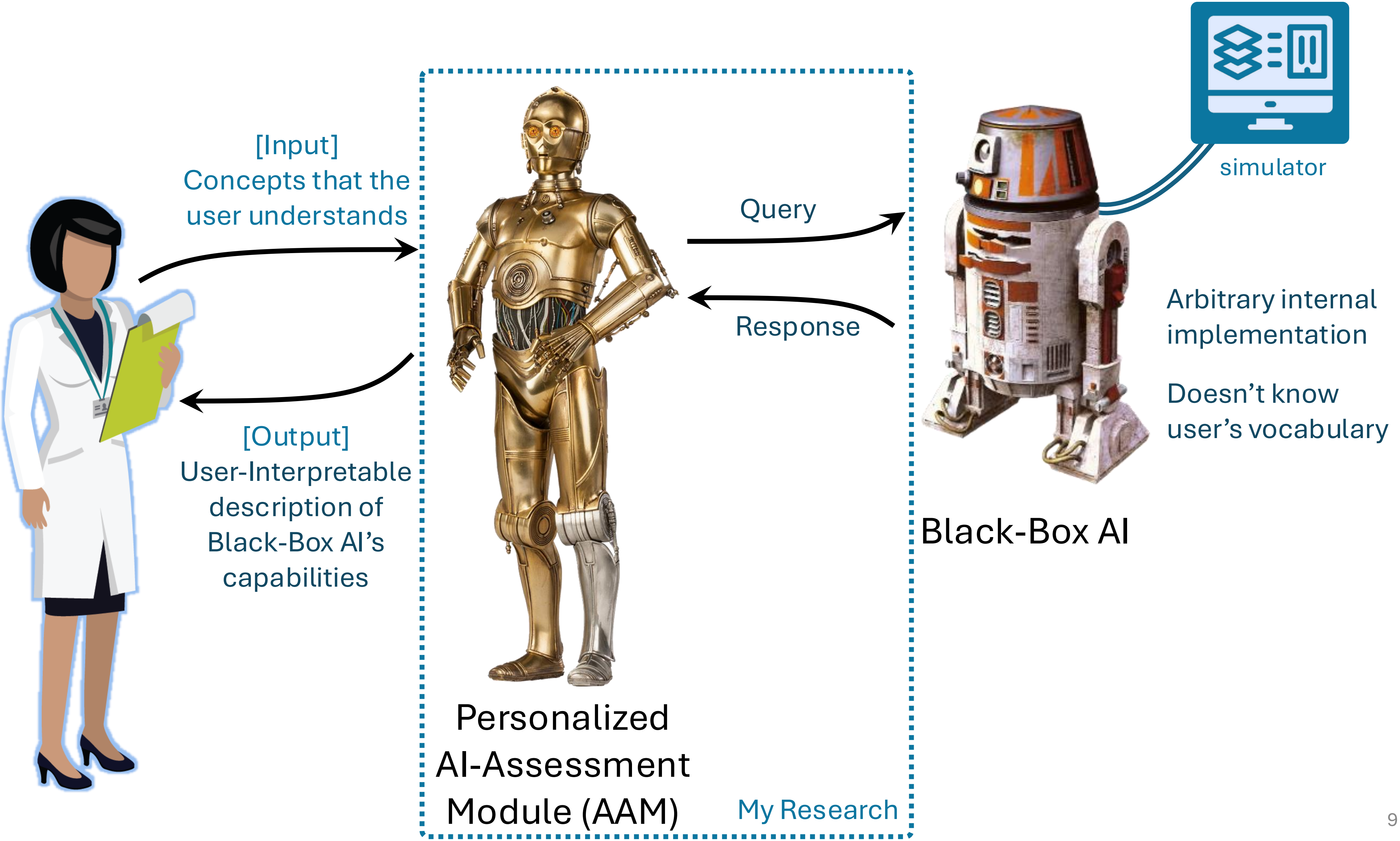
```
(:action open-door
   :parameters (?l1)
   :precondition (and
       (has_key)
       (player_at ?l1)
       (door_adjacent ?l1))
   :effect (probabilistic
       0.95 (and (door_open))
       0.05 (and (not(has_key))
                 (game-over))
)
```

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


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

Deterministic and Stationary Setting

Input

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

Output

- PDDL-like description of each capability.

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*)

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
    (+/-/∅)(has_key)
    (+/-/∅)(door_open)
    (+/-/∅)(door_adjacent ?l1)
    (+/-/∅)(player_at ?l1))
  :effect (and
    (+/-/∅)(has_key)
    (+/-/∅)(door_open)
    (+/-/∅)(door_adjacent ?l1)
    (+/-/∅)(player_at ?l1))
```

Stochastic and Stationary Setting

Input

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

Output

- PPDDL-like description of each capability.

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

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.

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.

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

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.

Capability Discovery and Assessment

Input

- Initial model of the AI system.
  - Predicates (User vocabulary)
    - With their evaluation functions

Output

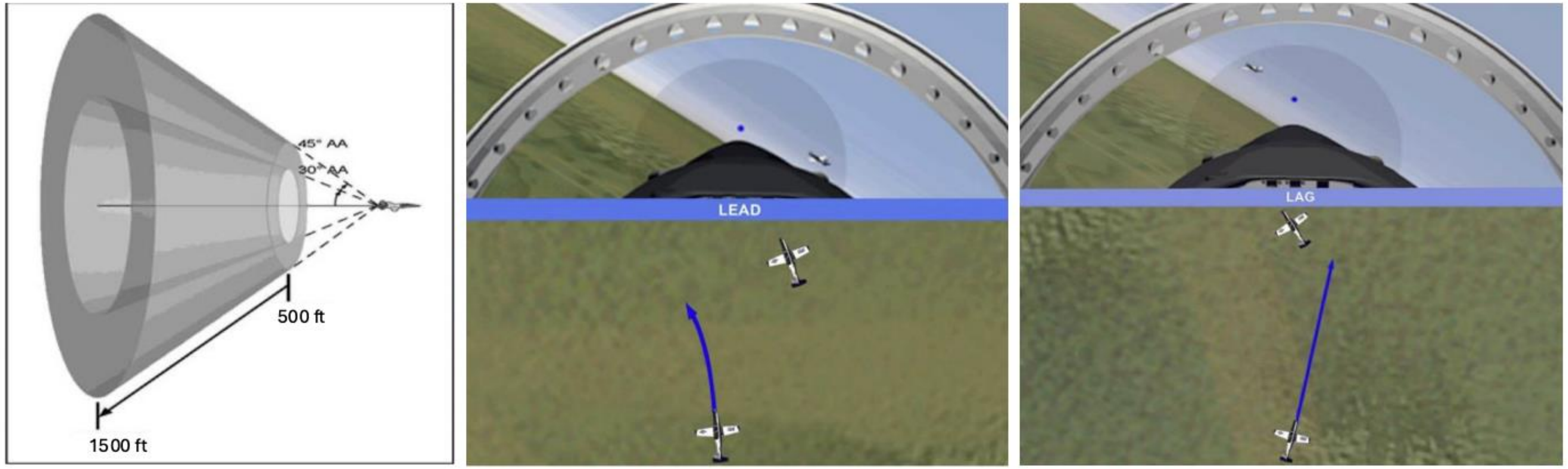
- List of capabilities
- PDDL-like description of each capability.

[Bramblett, Karia, Ciotinga, Suresh, **Verma**, Choi, Srivastava; ArXiv 2512.16733]

Assumptions

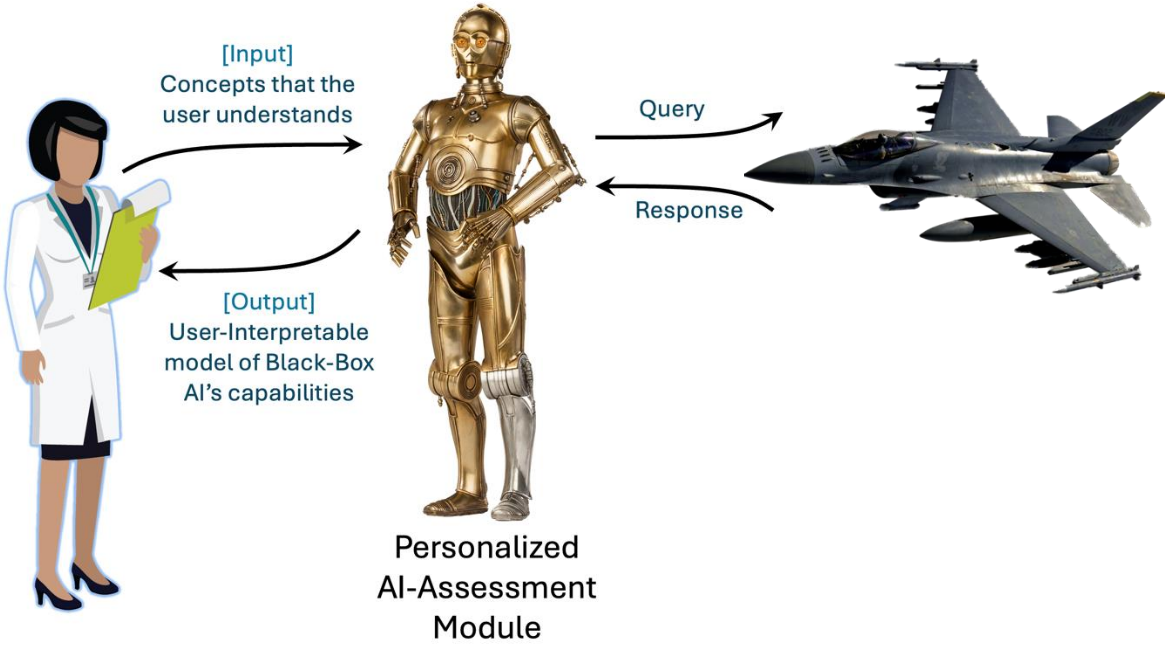
- ~~User's vocabulary matches simulator's vocabulary.~~
- Black-Box AI provides a list of ~~capabilities.~~ transitions.
- Stationary agent model.
- Stochastic ~~Deterministic~~ environment.
- Fully observable setting.

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?



[Rountree, **Verma**, So, Fan, Shah; openreview.net/forum?id=uEyJmixFiA]

Human Understanding of Deep Neural Policies

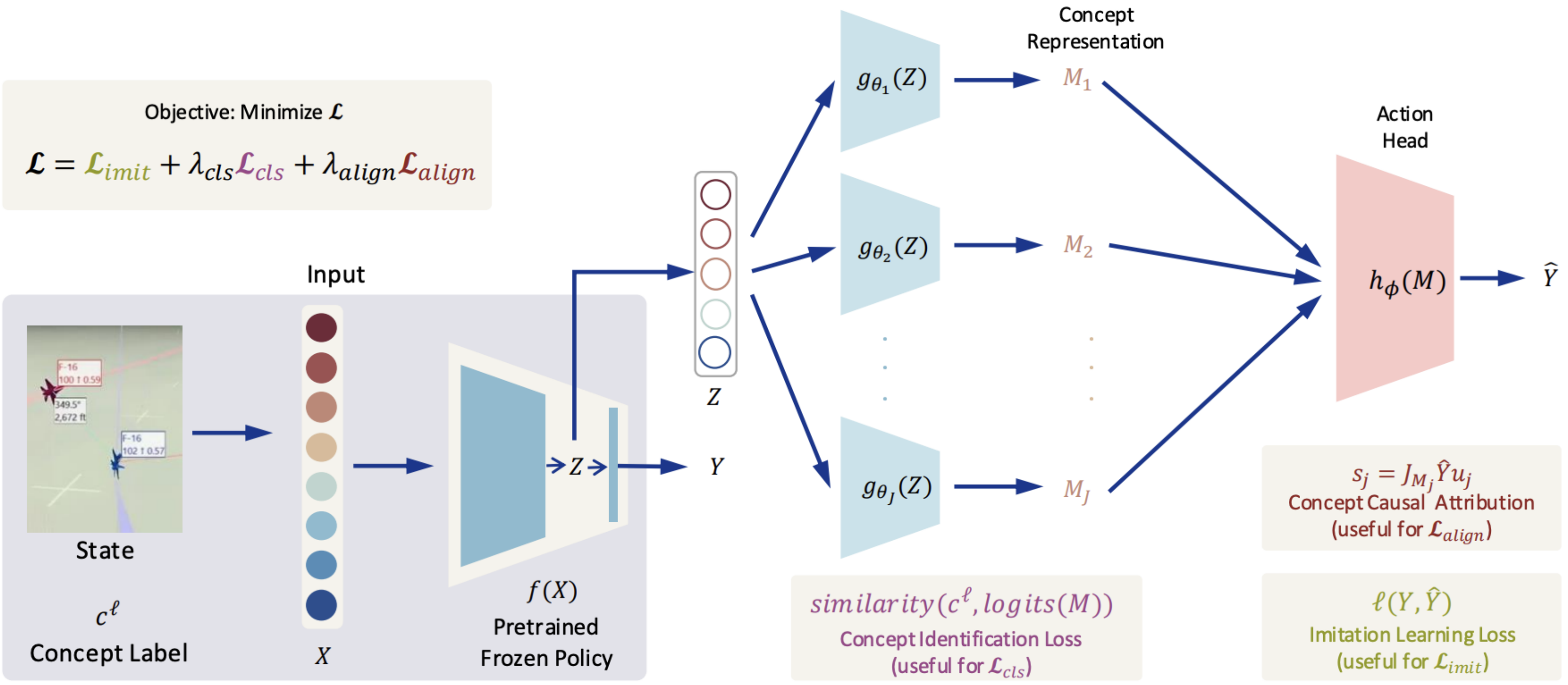


Figure 1: **CCW-Net Architecture.** CCW-Net wraps a frozen backbone f with a concept encoder g_θ that transforms latents Z into human-interpretable vector concepts M, and a differentiable head h_ϕ that maps concepts to wrapper actions $\hat{Y}$. CCW-Net estimates per-concept causal targets IE_j from expert trajectories, computes local concept-to-action sensitivities $s_j = J_{M_j}\hat{Y}\,u_j$, and aligns them with a masked cosine loss. Vector concepts encode both the strengths and context-dependent expression of each concept, enabling faithful, human-interpretable explanations while preserving task performance.

[Rountree, **Verma**, So, Fan, Shah; openreview.net/forum?id=uEyJmixFiA]

Lifelong Learning with Human-in-the Loop

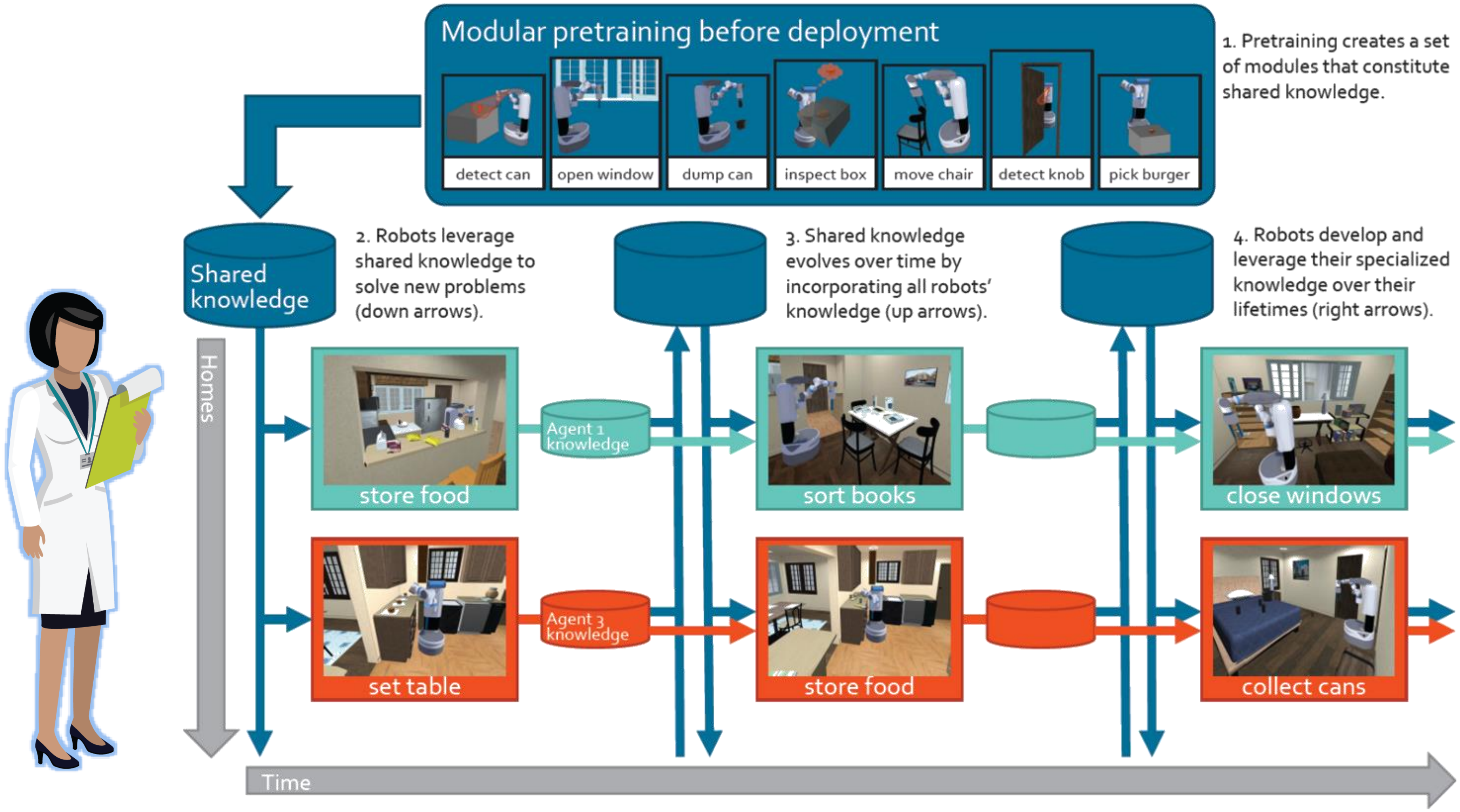
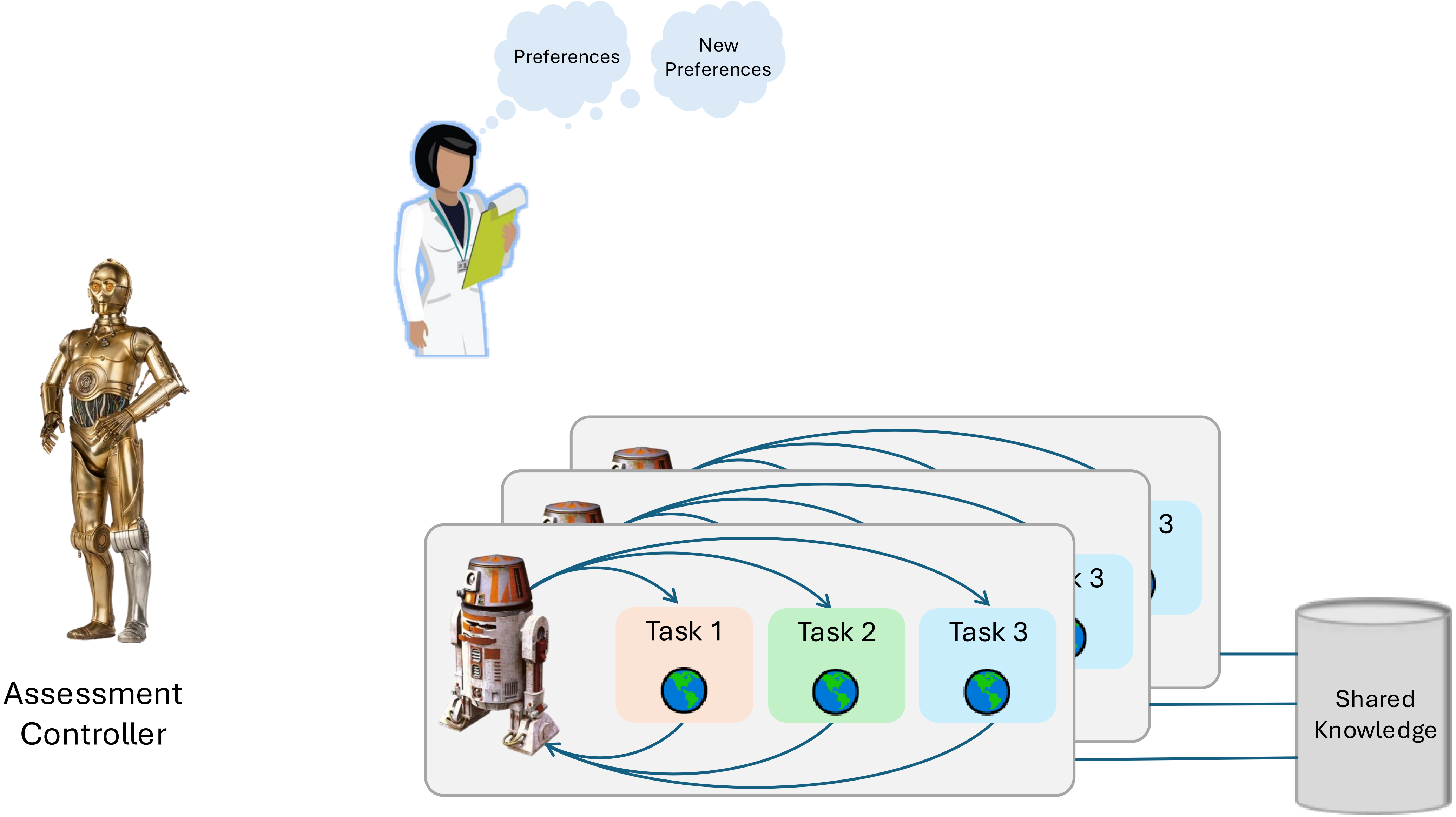


Image Courtesy: Jorge Mendez-Mendez, IRL Lab, MIT

Human-in-the Loop Lifelong Learning (HITL3)



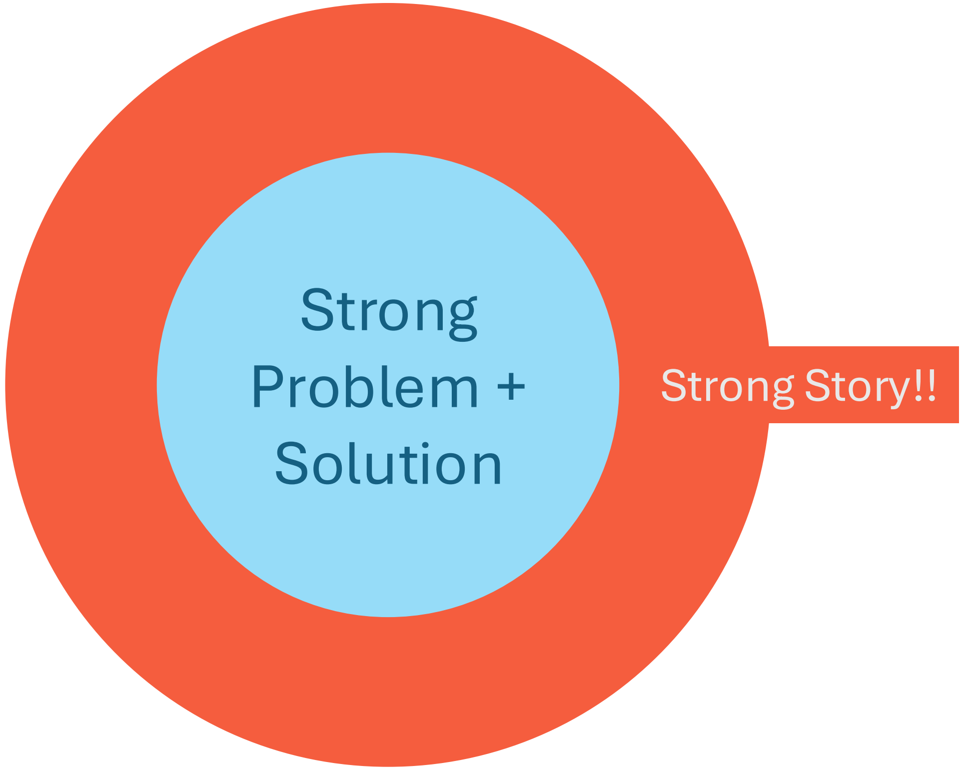
Some Advice for my future colleagues!!

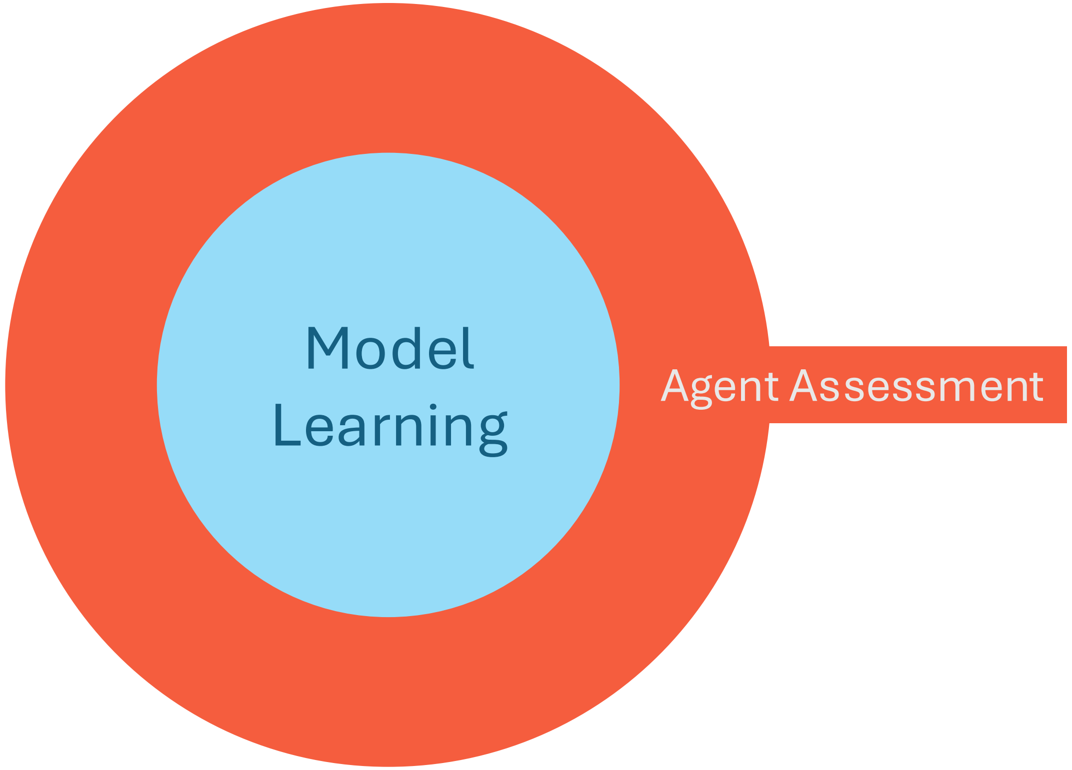
No hook, no fish

(aka No claims, No readers)*

*Subbarao Kambhampati. "Wittgenstein's papers & Faraday's Talks: Maxims for a milk-fed researcher". IJCAI 2013 DC

The Story Matters!!


Strong Problem + Solution
Strong Story!!


Model
Learning
Agent Assessment

Don’t go after the papers!!

Quality over Quantity

Take the reviews with a grain of salt!!

Rebuttal Matters!!

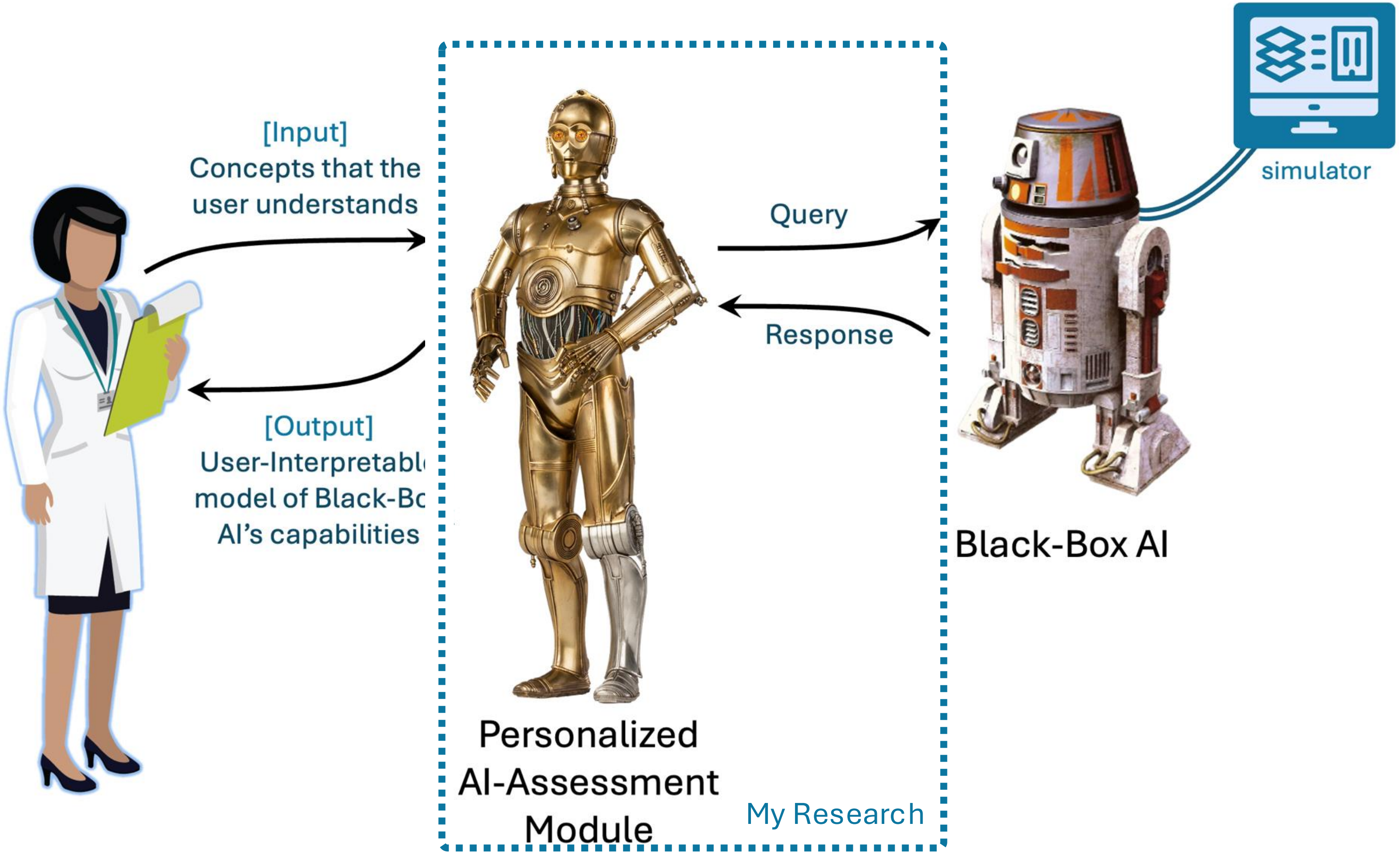
- Give your best
- Expect nothing

How to write rebuttals?

https://deviparikh.medium.com/how-we-write-rebuttals-dc84742fece1

Learning Interpretable Models for Autonomous Assessment of Taskable AI Systems

Pulkit Verma
pulkitv@iitm.ac.in



Questions?