



Learning Interpretable Models for Personalized Assessment of Taskable AI Systems

Pulkit Verma

✉ pulkitverma25@gmail.com

🌐 pulkitverma.net



01

Research Area &
Interesting Problems

02

Publications

03

Future Directions
at IIT Roorkee

04

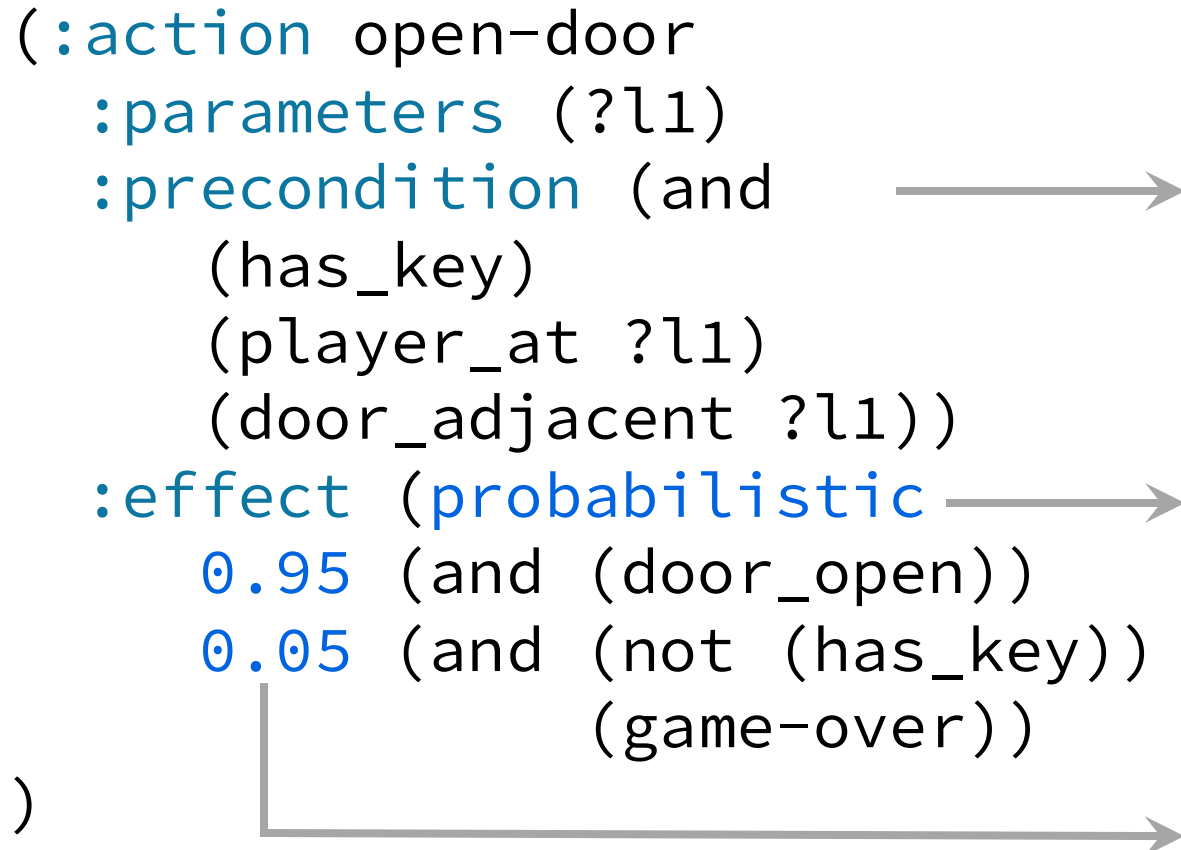
Teaching Plans
and Achievement

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



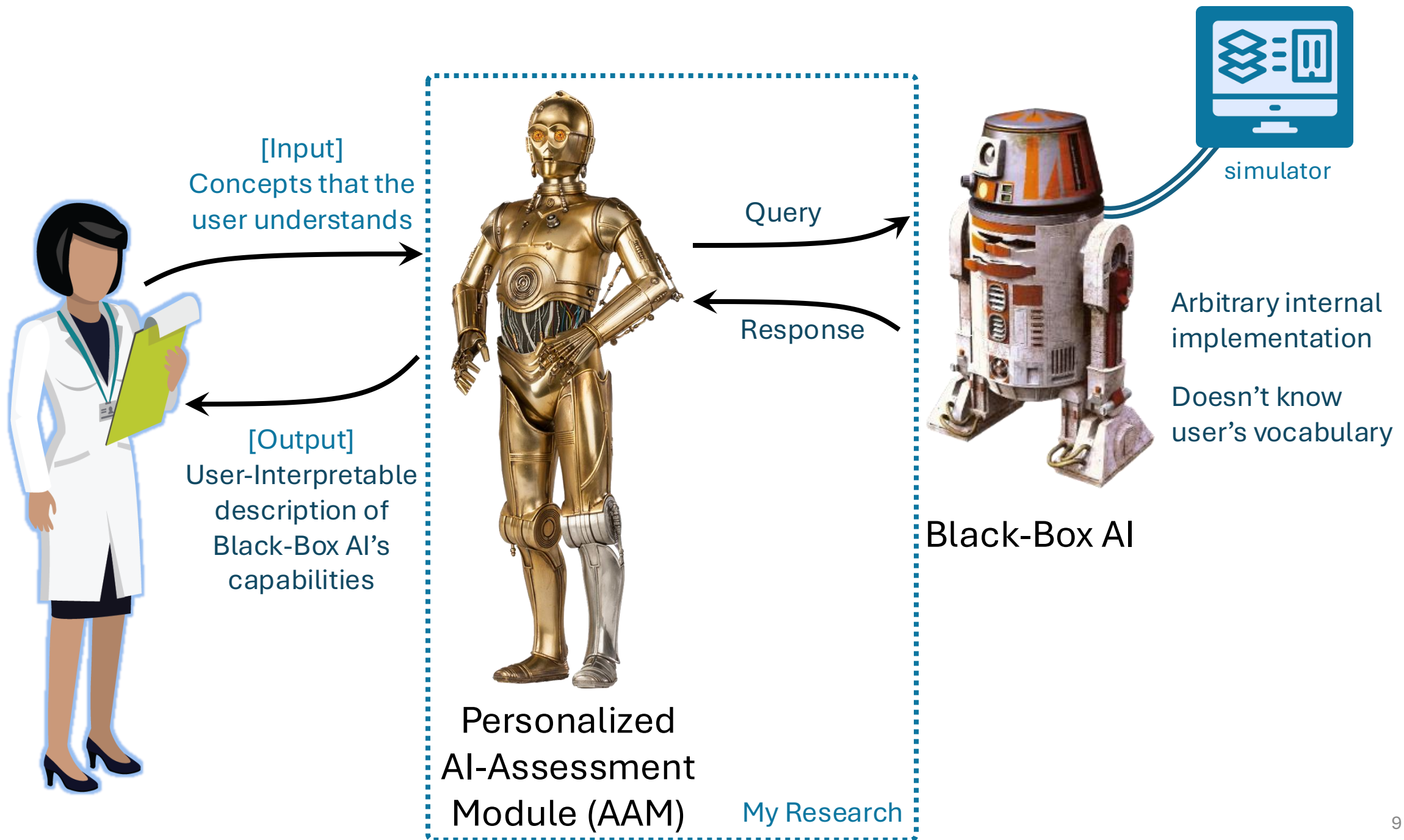
Black-Box AI



simulator

Arbitrary internal
implementation

Doesn't know
user's vocabulary



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.

Stochastic and Stationary Setting

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

[Verma, Karia, Srivastava; NeurIPS '23]

Discovering Capabilities

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

[Verma, Marpally, Srivastava; KR '22]

Differential Assessment

- 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]

01

Research Area &
Interesting Problems

02

Publications

03

Future Directions
at IIT Roorkee

04

Teaching Plans
and Achievement

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]

[Chen, Verma, Srivastava, Katz, Thiebaut; PRL@AAAI' 25]

Causal Accuracy of Symbolic Models

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

[**Verma**, Srivastava; Tech Report, 2024]

Explainable Robot Planning

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

[Dobhal, Nagpal, Karia, **Verma**, Nayyar, Shah, Srivastava; EAAI @ AAAI '25]

[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. A*
 - Rashmeet Kaur Nayyar*, **Pulkit Verma***, and Siddharth Srivastava. *Differential Assessment of Black-Box AI Agents*. In AAAI 2022. A*
 - **Pulkit Verma**, Shashank Rao Marpally, and Siddharth Srivastava. *Discovering User-Interpretable Capabilities of Black-Box Planning Agents*. In KR 2022. A*
 - Naman Shah*, **Pulkit Verma***, Trevor Angle, and Siddharth Srivastava. *JEDAI: A System for Skill-Aligned Explainable Robot Planning*. In AAMAS 2022 (Demonstration Track). A*
-  Winner of Best Demo Award
- Yizhong Wang et al. *Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ Tasks*. In EMNLP 2022. A
 - **Pulkit Verma**, Rushang Karia, and Siddharth Srivastava. *Autonomous Capability Assessment of Sequential Decision-Making Systems in Stochastic Settings*. In NeurIPS 2023. A*
 - Rushang Karia*, **Pulkit Verma***, Alberto Speranzon, and Siddharth Srivastava. *Epistemic Exploration for Generalizable Planning and Learning in Non-Stationary Settings*. In ICAPS 2024. A*
 - Rushang Karia*, Jayesh Nagpal*, Daksh Dobhal*, **Pulkit Verma**, Rashmeet Kaur Nayyar, Naman Shah, and Siddharth Srivastava. *Using Explainable AI and Hierarchical Planning for Outreach with Robots*. In EAAI 2025 (Proceedings of AAAI'25). A*

*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.
- Dillon Z. Chen, **Pulkit Verma**, Siddharth Srivastava, Michael Katz, and Sylvie Thiébaux. *AI Planning: A Primer and Survey (Preliminary Report)*. In AAAI 2025 PRL.
- 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.
- 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.

01

Research Area &
Interesting Problems

02

Publications

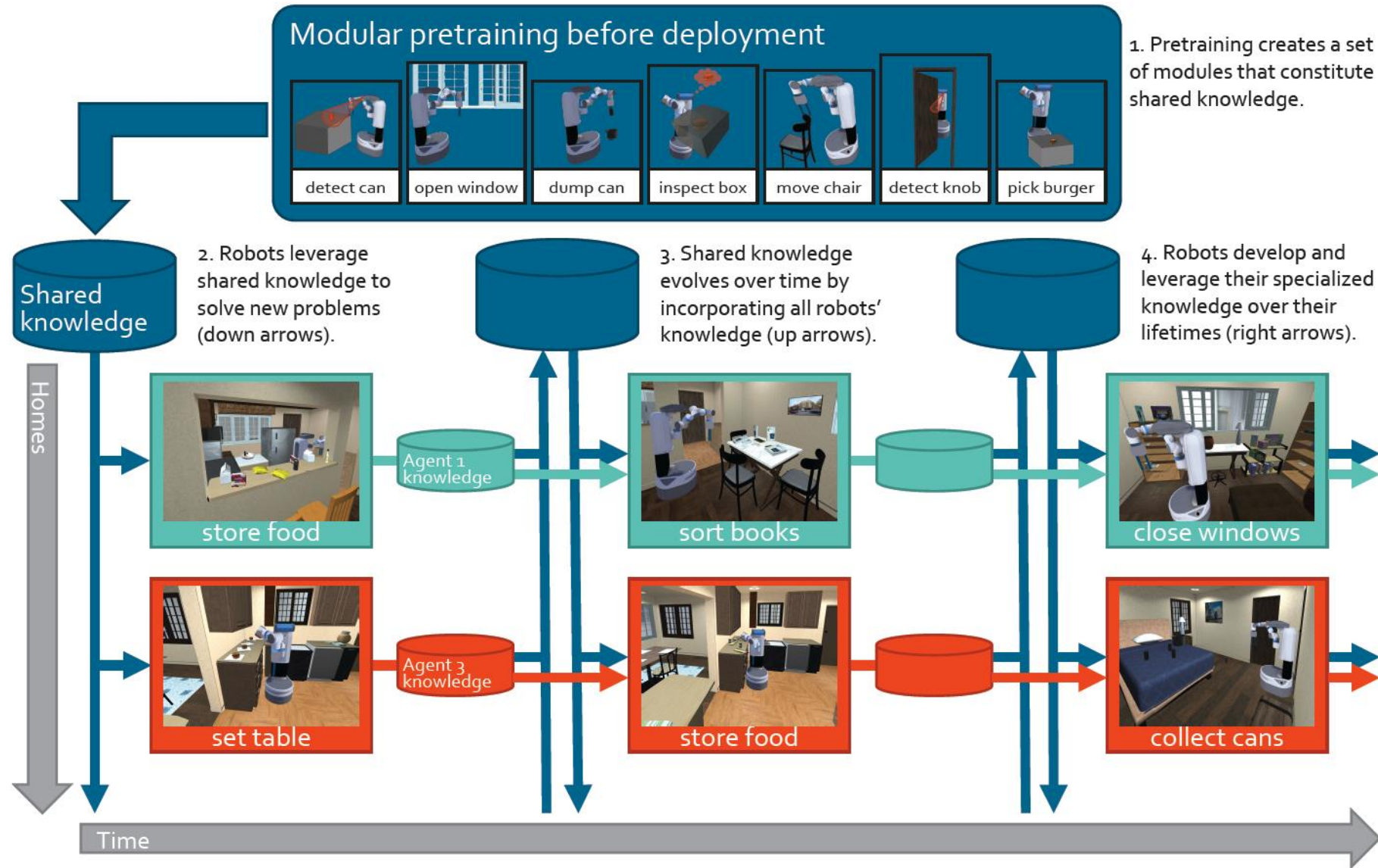
03

Future Directions
at IIT Roorkee

04

Teaching Plans
and Achievement

Lifelong Learning for Embodied Agents



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

01

Research Area &
Interesting Problems

02

Publications

03

Future Directions
at IIT Roorkee

04

Teaching Plans
and Achievement

Teaching Plans

AI Courses



Awarded Teaching Excellence Award for this Course at ASU

In addition to the existing syllabus, I plan to add ROS based experiments to give practical experience to students.

Explore the symbolic planning part of AI, in addition to integrating it with reinforcement learning based approaches.

Classic logic and knowledge representation, answer set programming, reasoning about actions and planning, etc.
Long term goal - probably in 3-4 years

Integrating paper-reading with some introductory lectures on topics required for understanding the papers.
AI Course will be a prerequisite

Suitable for M. Tech. and
Final year B.Tech. students

Artificial Intelligence

Planning and Learning
Methods in AI

Knowledge Representation
and Reasoning

For
PhDs

AI Safety and Assessment

Teaching Plans

AI Courses

Suitable for M. Tech. and
Final year B.Tech. students

Artificial Intelligence

Planning and Learning
Methods in AI

Knowledge Representation
and Reasoning

For
PhDs

AI Safety and Assessment

Non-AI Courses

Algorithms and Data Structures

Design and Analysis of Algorithms

Operating Systems

Database Management System

Learning Interpretable Models for Personalized Assessment of Taskable AI Systems

Pulkit Verma

pulkitverma25@gmail.com

