AAAI 2024 Spring Symposium

User-Aligned Autonomous Capability Assessment of Black-Box AI Systems

Pulkit Verma and ***Siddharth Srivastava*** | *Arizona State University*

AAI Autonomous Agents and Intelligent Robots

ASU Arizona State University

How would a non-expert assess the limits and capabilities of an AI system?

Objective

Learn an interpretable model of an adaptive taskable AI system by interrogating it.

Approach

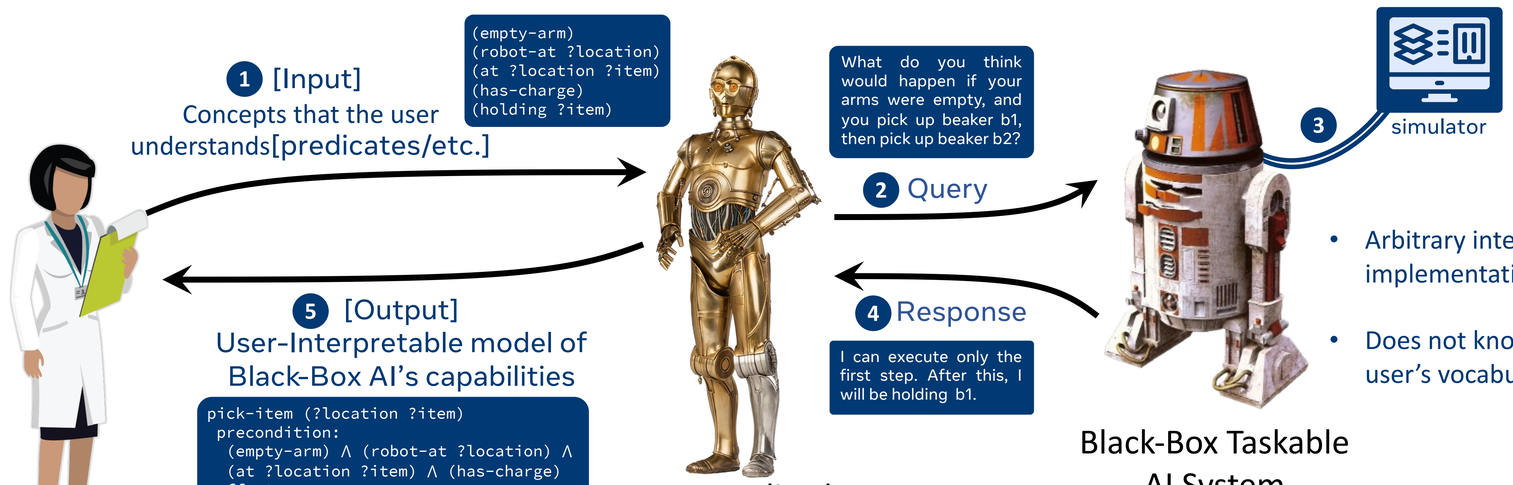
- Create an interface and a minimal set of requirements in an AI system that would enable their assessment using this interface.

- Learn an *interpretable* model of a taskable sequential decision-making AI system.

Summary

- Efficiently learns the model of a taskable AI system in a STRIPS-like form.
- Needs no prior knowledge of the AI system's model.
- Only requires an AI system to have rudimentary query answering capabilities.
- Queries can be answered using a simulator.

Personalized AI Assessment Framework

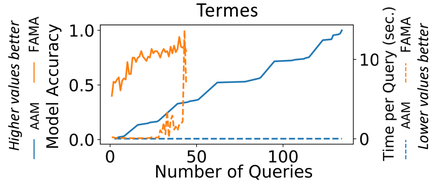


- Need not be an AI expert
- Needs to understand the AI system's limits and capabilities
- Arbitrary internal implementation
- Does not know the user's vocabulary

Desiderata

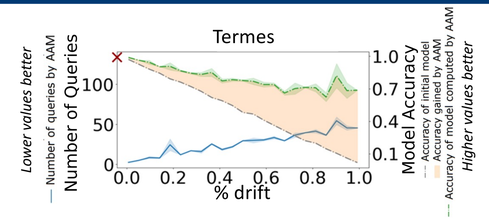
- ***Interpretable Description***
  Users should be able to understand the description.
- ***Correct Description***
  The generated model should be accurate.
- ***Generalizable Design***
  AAM should work for a variety of Taskable AI Systems.
- ***Easy to Satisfy Requirements***
  The requirements for the AI system to support the assessment should be easy to support for greater adoption.

Results



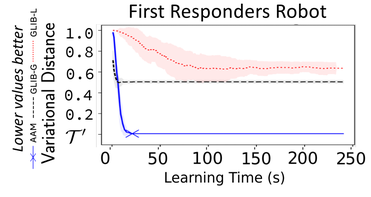
Verma, Marpally, Srivastava. *Asking the Right Questions: Learning Interpretable Action Models Using Query Answering.* AAAI 2021.

AAM always learns an accurate model faster compared to passive learners (FAMA).



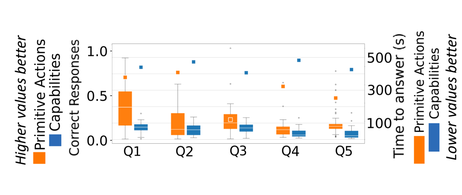
Nayyar*, **Verma***, Srivastava. *Differential Assessment of Black-Box AI Systems.* AAAI 2022.

Learning a model's drifted parts is much faster than learning the whole model from scratch.



Verma, Karia, Srivastava. *Autonomous Capability Assessment of Sequential Decision-Making Systems in Stochastic Settings.* NeurIPS 2023.

AAM can learn a probabilistic model closer to the true model than state-of-the-art.



Verma, Marpally, Srivastava. *Discovering User-Interpretable Capabilities of Black-Box Planning Agents.* KR 2022.

AAM discovers interpretable high-level capabilities that users can use to reason with correctly.