User Driven Assessment of Adaptive Taskable AI Systems

Pulkit Verma | AAIR Lab, Arizona State University | Advised by: Siddharth Srivastava

Center for Human-Compatible Artificial Intelligence

Autonomous Agents and Intelligent Robots — Arizona State University

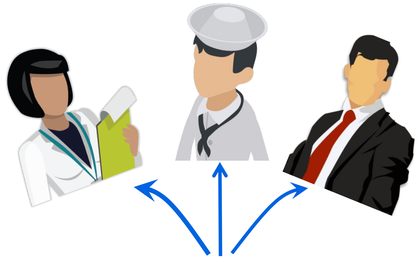
How would a non-expert assess the limits and capabilities of an AI system?

Introduction

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.

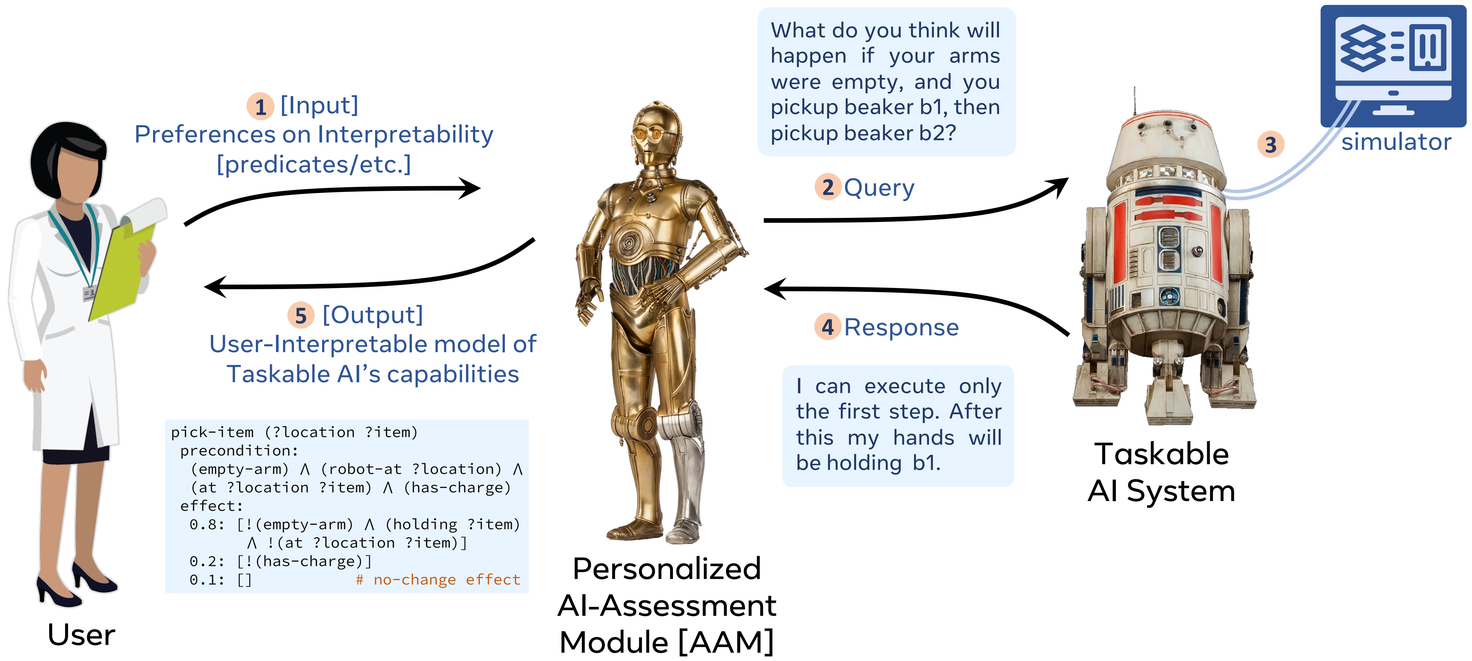


- Learn an interpretable model of a taskable sequential decision-making AI system.

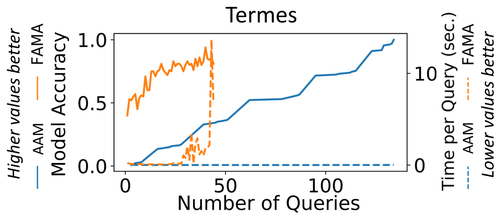
Summary

- 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.

Agent Assessment Framework

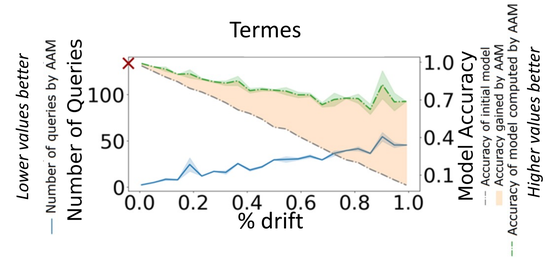


Results



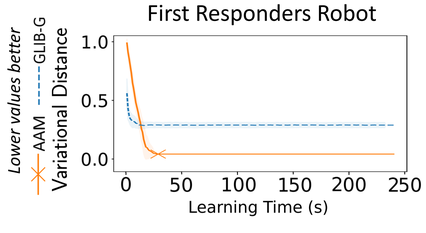
Verma, Marpally, Srivastava. *Asking the Right Questions: Learning Interpretable Action Models Using Query Answering.* AAAI 2021.

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



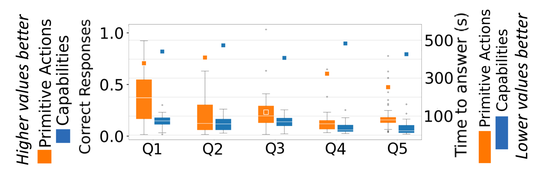
Nayyar*, **Verma***, Srivastava. *Differential Assessment of Black-Box AI Systems.* AAAI 2022.

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



Verma, Karia, Srivastava. *Autonomous Capability Assessment of Black-Box Sequential Decision-Making Systems.* ICAPS 2023 KEPS Workshop.

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.