# Data Efficient Paradigms for Personalized Assessment of Black-Box Taskable AI Systems

**Pulkit Verma**

Autonomous Agents and Intelligent Robots Lab,
School of Computing and Augmented Intelligence, Arizona State University, USA
verma.pulkit@asu.edu

## Abstract

The vast diversity of internal designs of taskable black-box AI systems and their nuanced zones of safe functionality make it difficult for a layperson to use them without unintended side effects. My dissertation focuses on developing paradigms that enable a user to assess and understand the limits of an AI system's safe operability. We develop a personalized AI assessment module that lets an AI system execute instruction sequences in simulators and answer queries about these executions. Our results show that such a primitive query-response interface is sufficient to efficiently derive a user-interpretable model of a system's capabilities.

## Introduction

The growing deployment of AI systems presents a pervasive problem of ensuring the safety and reliability. To use them safely, the lay people using black-box AI systems need to understand how they work or what they can and cannot do. We develop a paradigm that allow for assessment of such black-box AI systems in terms of their capabilities. We also define the set of requirements in terms of a minimal query-response interface that the black box AI should support for such an assessment. This paradigm can also be extended in settings where an AI system's capabilities are evolving and/or adapting to changes in the environment it is working in. In such dynamic settings, such a solution can help a lay user ensure the reliable and safe usage of the AI system. Our assessment approach generates a description of the capabilities of the AI system so that the non-experts can understand their limits and capabilities. This is important as lack of such knowledge can lead to unsafe usage, or serious accidents.

## Focus of My Dissertation

In my dissertation, I plan to develop a *personalized AI-assessment module* (AAM), shown in Fig. 1, which can derive the model of capabilities of a black-box AI system in terms of an user-interpretable vocabulary. AAM takes as input using as input (i) the agent (ii) a compatible simulator using which the agent can simulate its primitive action sequences; and (iii) the user's concept vocabulary, which


Figure 1: The personalized AI assessment module uses the user's preferred vocabulary, queries the AI system, and delivers an interpretable model of the AI system's capabilities.

may be insufficient to express the simulator's state representation. AAM queries the AI system and receives its responses. At the end of the querying process, AAM returns a user-interpretable model of the AI system's capabilities. This approach's advantage is that the AI system need not know the user vocabulary or the modeling language. In my work's context, "actions" refer to the core *functionality* of the agent, denoting the agent's decision choices or primitive actions that the agent could execute. In contrast, "capabilities" refer to the *high-level behaviors* that the AI system can perform using its behavior synthesis algorithms.

**Generating Interrogation Policies** I aim to create an interrogation policy that will generate queries for the AI system, and use the answers to estimate its model in the user-interpretable vocabulary. I plan to generate these queries by reducing the query generation to a planning problem and then use an interrogation algorithm to iteratively generate new queries, based on responses to previous queries.

**Inferring the Capability Model** Given the predicates and capabilities, there is an exponential number of PDDL models possible. To avoid this combinatorial explosion, I plan to use a top-down process that eliminates large classes of models, inconsistent with the AI system, by computing queries that discriminate between pairs of *abstract models*. When an abstract model's answer to a query differs from that of the AI system, we can eliminate the entire set of possible models that are refinements of this abstract model. I plan to start research on this with simplistic queries in fully observable environments and expand the scope to more general settings.

**Discovering the Capabilities and Learning their Descriptions** I want the assessment module to discover the high-level capabilities of a taskable AI system. I plan to collect a set of observations capturing the AI system's behavior in form of the state transitions. I would then discover the high-level capabilities of the AI system's behavior using those state transitions, and learn these capabilities' description.

## Related Work

Several action model learning approaches (Arora et al. 2018) have focused on learning the AI system's model using passively observed data. These approaches do not feature any interventions, hence are susceptible to learning buggy models. Unlike these, our approach queries the AI system and is guaranteed to converge to the true model while presenting a running estimate of the learned model's accuracy; so, it can be used in settings where the AI system's model can change.

## Preliminary Results

We developed four preliminary versions of the personalized AI assessment module, which we discuss briefly below.

**Learning the capability model** The first preliminary version of the AI assessment module, called the agent interrogation algorithm (AIA) (Verma, Marpally, and Srivastava 2021), efficiently derives a user-interpretable model of the system in stationary, fully observable, and deterministic settings. We compared AIA with the closest related work FAMA (Aineto, Celorrio, and Onaindia 2019) and found that AIA takes lesser time per query and shows better convergence to correct model. We also show that the models that we learn capture the correct causal relationships in the AI system's behavior in terms of how the system operates and interacts with its environment (Verma and Srivastava 2021), unlike the approaches that only use observations.

**Differential assessment** We developed a *differential assessment* version of the personalized AI assessment module, called DAAISy (Nayyar, Verma, and Srivastava 2022). This addresses the problem of accurately predicting the behavior of a black-box AI system that is evolving and adapting to changes in the environment it is operating in. DAAISy utilizes an initially known PDDL model of the AI system in the past, and a small set of observations of AI system's execution. It uses these observations to develop a querying strategy that avoids the full cost of assessment from scratch and outputs a revised model of the system's new functionality.

**Discovering the capabilities and learning their descriptions** We also developed a version of AAM that can discover high-level capabilities of an AI planning agent expressible in terms of the user-interpretable concept vocabularies (Verma, Marpally, and Srivastava 2022). The descriptions of these capabilities as a model are returned to the user as opposed to the model of the agent's primitive actions. We also conducted a user study to evaluate interpretablity of the capability descriptions computed by our approach. The results of this study showed that the users take less time to answer questions and they got more responses correct when using the capabilities as compared to using primitive actions.

**Learning a probabilistic capability model** We also created a version of AAM, called the query-based autonomous capability estimation (QACE) (Verma, Karia, and Srivastava 2023), that efficiently derives a user-interpretable model of the system's capabilities in stochastic settings. We compared QACE with the closest related work GLIB (Chitnis et al. 2021) in terms of the learned model's accuracy and the time taken to learn the model. We found that it leads to few shot generalization, convergence to a sound and complete model, and greater sample efficiency and accuracy for learning lifted relational models for AI systems compared to the baseline. We also extend it to discover a model of the agent's stochastic capabilities (Verma et al. 2023), and to make RL sample efficient (Karia et al. 2023).

## Conclusions and Future Work

This work presents a novel framework for learning the capability description of an AI system in terms of user-interpretable concepts by combining information from passive execution traces and active query answering. This approach also works for settings where the user's conceptual vocabulary is imprecise and cannot directly express the agent's capabilities. Next, I plan to use it with systems like JEDAI (Shah et al. 2022) as interfaces to make AI systems compliant with Level II assistive AI (Srivastava 2021).

## References

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

Arora, A.; Fiorino, H.; Pellier, D.; Métivier, M.; and Pesty, S. 2018. A Review of Learning Planning Action Models. *Knowledge Engineering Review*, 33: E20.

Chitnis, R.; Silver, T.; Tenenbaum, J.; Kaelbling, L. P.; and Lozano-Pérez, T. 2021. GLIB: Efficient Exploration for Relational Model-Based Reinforcement Learning via Goal-Literal Babbling. In *Proc. AAAI*.

Karia, R.; Verma, P.; Vipat, G.; and Srivastava, S. 2023. Epistemic Exploration for Generalizable Planning and Learning in Non-Stationary Stochastic Settings. In *NeurIPS 2023 GenPlan Workshop*.

Nayyar, R. K.; Verma, P.; and Srivastava, S. 2022. Differential Assessment of Black-Box AI Agents. In *Proc. AAAI*.

Shah, N.; Verma, P.; Angle, T.; and Srivastava, S. 2022. JEDAI: A System for Skill-Aligned Explainable Robot Planning. In *Proc. AAMAS*.

Srivastava, S. 2021. Unifying Principles and Metrics for Safe and Assistive AI. In *Proc. AAAI*.

Verma, P.; Karia, R.; and Srivastava, S. 2023. Autonomous Capability Assessment of Sequential Decision-Making Systems in Stochastic Settings. In *Proc. NeurIPS*.

Verma, P.; Karia, R.; Vipat, G.; Gupta, A.; and Srivastava, S. 2023. Learning AI-System Capabilities under Stochasticity. In *NeurIPS 2023 GenPlan Workshop*.

Verma, P.; Marpally, S. R.; and Srivastava, S. 2021. Asking the Right Questions: Learning Interpretable Action Models Through Query Answering. In *Proc. AAAI*.

Verma, P.; Marpally, S. R.; and Srivastava, S. 2022. Discovering User-Interpretable Capabilities of Black-Box Planning Agents. In *Proc. KR*.

Verma, P.; and Srivastava, S. 2021. Learning Causal Models of Autonomous Agents using Interventions. In *IJCAI 2021 GenPlan Workshop*.