Sample Efficient Paradigms for Personalized Assessment of 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. The focus of my dissertation is to develop paradigms that would 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 capability is sufficient to efficiently derive a user-interpretable model of the system's capabilities in fully observable settings.

1 Introduction

The growing deployment of AI systems presents a pervasive problem of ensuring the safety and reliability of these systems. The problem is exacerbated because most of these AI systems are neither designed by their users nor are their users skilled enough to understand their internal working, i.e., the AI system is a black-box for them. Hence such systems may be used by non-experts who may not understand how they work or what they can and cannot do. Ongoing research on the topic focuses on the significant problem of answering such a user's questions about the system's behavior [Chakraborti *et al.*, 2017]. However, most non-experts do not know which questions to ask for assessing the safe limits and capabilities of an AI system. This problem is aggravated in situations where an AI system can carry out planning or sequential decision making. My dissertation work aims to create general algorithms and methods for interpretability which when used with a black-box AI system, can help generate a description of its capabilities by interrogating it.

2 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 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 the context of my work, "actions" refer to the core *functionality* of the agent, denoting the agent's decision choices or primitive actions that the agent could execute (e.g., keystrokes in a video game). In contrast, "capabilities" refer to the *high-level behaviors* that the AI system can perform using its AI algorithms for behavior synthesis, including planning and learning (e.g., navigating to a room).

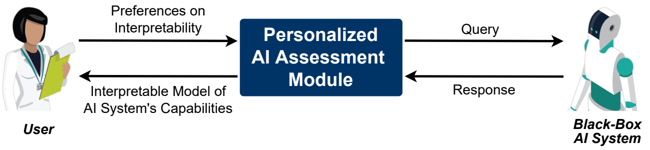
Preferences on Interpretability
Personalized AI Assessment Module
Query
Interpretable Model of AI System's Capabilities
Response
User
Black-Box AI System

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.

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 Action Model Given the predicates and actions, 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 front with simplistic queries in fully observable environments and expand the scope to more general settings. In the

[*]Advisor: **Siddharth Srivastava**, Arizona State University

future, this mechanism can be extended to complex queries.

Discovering the Capabilities and Learning their Descriptions I want the assessment module to discover the high-level capabilities of the AI system that can plan, and not just the action model of an AI system. I plan to collect a set of state observations capturing the behavior of the AI system 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 the description of these capabilities.

3 Related Work

Several action model learning approaches [Arora *et al.*, 2018; Aineto *et al.*, 2019] 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 approaches, our approach queries the AI system and is guaranteed to converge to the true model while presenting a running estimate of the accuracy of the derived model; hence, it can be used in settings where the AI system's model changes over time.

4 Preliminary Results

We developed four preliminary versions of the personalized AI assessment module, each focusing on one specific subproblem of the overall larger goal.

Learning the action model The first preliminary version of the AI assessment module, called the agent interrogation algorithm (AIA) [Verma and Srivastava, 2020; Verma *et al.*, 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 *et al.*, 2019] in terms of the learned model's accuracy, the number of queries asked, and the time taken to generate those queries. For systems initialized with IPC domains, AIA takes lesser time per query and shows better convergence to the 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 observational data.

Differential assessment We developed a *differential assessment* version of the personalized AI assessment module, called DAAISy [Nayyar *et al.*, 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 an incremental 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 *et al.*, 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 the behavior analysis 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 action model We also created a version of AAM, called the query-based autonomous capability estimation (QACE) [Verma *et al.*, 2023], that efficiently derives a user-interpretable model of the system's actions 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 QACE leads to (i) few shot generalization; (ii) convergence to a sound and complete model; and (iii) much greater sample efficiency and accuracy for learning lifted relational models for AI systems with complex capabilities as compared to the baseline.

I plan to extend it to learn a model of the agent's capabilities in partially observable settings, and 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 *et al.*, 2019] D. Aineto, S. J. Celorrio, and E. Onaindia. Learning action models with minimal observability. *Artificial Intelligence*, 275:104–137, 2019.

[Arora *et al.*, 2018] A. Arora, H. Fiorino, D. Pellier, M. Métivier, and S. Pesty. A review of learning planning action models. *Knowledge Engineering Review*, 33:E20, 2018.

[Chakraborti *et al.*, 2017] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*, 2017.

[Chitnis *et al.*, 2021] R. Chitnis, T. Silver, J. Tenenbaum, L. P. Kaelbling, and T. Lozano-Pérez. GLIB: Efficient exploration for relational model-based reinforcement learning via goal-literal babbling. In *AAAI*, 2021.

[Nayyar *et al.*, 2022] R. K. Nayyar, P. Verma, and S. Srivastava. Differential assessment of black-box AI agents. In *AAAI*, 2022.

[Shah *et al.*, 2022] N. Shah, P. Verma, T. Angle, and S. Srivastava. JEDAI: A system for skill-aligned explainable robot planning. In *AAMAS*, 2022.

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

[Verma and Srivastava, 2020] P. Verma and S. Srivastava. Learning generalized models by interrogating black-box autonomous agents. In *AAAI 2020 GenPlan Workshop*, 2020.

[Verma and Srivastava, 2021] P. Verma and S. Srivastava. Learning causal models of autonomous agents using interventions. In *IJCAI 2021 GenPlan Workshop*, 2021.

[Verma *et al.*, 2021] P. Verma, S. R. Marpally, and S. Srivastava. Asking the right questions: Learning interpretable action models through query answering. In *AAAI*, 2021.

[Verma *et al.*, 2022] P. Verma, S. R. Marpally, and S. Srivastava. Discovering user-interpretable capabilities of black-box planning agents. In *KR*, 2022.

[Verma *et al.*, 2023] P. Verma, R. Karia, and S. Srivastava. Autonomous capability assessment of black-box sequential decision-making systems. In *ICAPS 2023 KEPS Workshop*, 2023.