Learning Causal Models of Autonomous Agents using Interventions

Pulkit Verma, Siddharth Srivastava | Arizona State University

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

Introduction

Objective: Learn an interpretable model of a black-box agent by interrogating it.

Key technical challenge:
- Which sequence of queries to ask?

Causal Models

Definition 1. A *causal model* M is defined as a 4-tuple $\langle \mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F} \rangle$ where $\mathcal{U}$ is a set of exogenous variables (representing factors outside the model's control), $\mathcal{V}$ is a set of endogenous variables (whose values are directly or indirectly derived from the exogenous variables), $\mathcal{R}$ is a function that associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y, and $\mathcal{F}$ is a function that associates with each endogenous variable $X \in \mathcal{V}$ a structural function denoted as F_X such that F_X maps $\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)$ to $\mathcal{R}(X)$.

Definition 2. $\vec{X} = \vec{x}$ is an *actual cause* of φ in the causal setting $(M, \vec{u})$ if the following conditions hold:

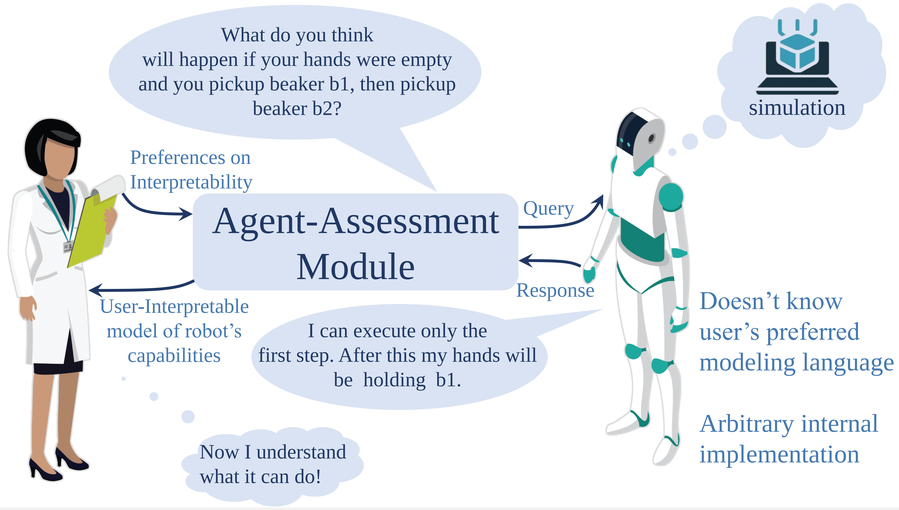
AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$.

AC2. There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*] \neg\varphi$.

AC3. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies conditions AC1 and AC2, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.

Lemma: The action models learned by the agent interrogation algorithm are causal models.

Example of Agent Interrogation



Comparing Queries

- How difficult is it to evaluate/use query responses.
- How difficult is it to answer a query.

Plan Outcome Queries
- Contains Initial State and Plan
- Return length of successful execution and final state
- Difficult to learn model from query responses.
- Easy to answer*.

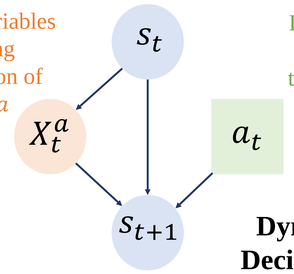
Action Precondition Queries
- Contains Initial State and Plan
- Return length of successful execution and failing preconditions
- Easy to learn model from query responses
- Difficult to answer*.

Salient Features

- Efficiently learns **causal model** of an AI agent in STRIPS-like form.
- Needs no prior knowledge of the agent model.
- Only requires an agent to have rudimentary query answering capabilities.

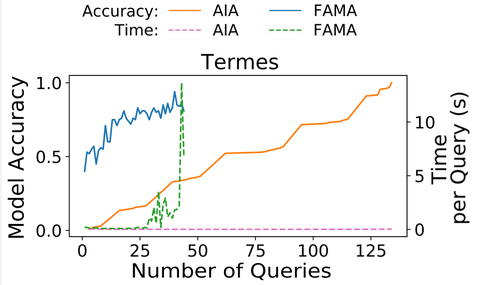
Dynamic Causal Decision Networks

- **Hard Interventions**: Searching for initial states in AIA. Helps in learning preconditions of an action.
- **Soft Interventions**: Setting values for decision nodes. Helps in learning effects of an action.



Dynamic Causal Decision Network

Results



- AIA efficiently derives interpretable agent models for a range of agents.
- AIA is much faster than state of the art methods for deriving models based on passive observations.
- AIA offers better convergence guarantees.

Complexity Results

Complexity of learning the action model based on responses:
- Plan Outcome Queries: O(|P|x|A|).
- Action Precondition Queries: O(|A|).

Membership classes for both plan outcome and action precondition queries
- Data Complexity: AC^0
- Expression Complexity: ALOGTIME
- Combined Complexity: PTIME

Theorem: Given an agent ***A*** with an unknown ground truth model $\boldsymbol{M^A}$, the action model ***M*** learned by the agent interrogation algorithm is **sound and complete.**

It is not necessary that a method using only observations learns models that are sound or complete.

Refer to the papers for detailed results

bit.ly/3p4cVRu

bit.ly/3eNcW9G