HP-Causality in Multi-agent Planning Tasks

Axel Ind

November 29, 2018

Abstract

HP-Causality (@TODO reference) allows us to determine which variable values are causes of certain other final variable values. The normal HP-Causality approach is limited to a structural equation representations of the problem under consideration. This paper describes a multi-agent planning approach to defining causality. Particular focus is paid to the Modified definition of HP-causality.

1 Introduction

2 Preliminaries

2.1 Single-Agent Planning Tasks

A single-agent planning task is a 4-tuple $\Pi =< V, I, O, \gamma >$. Where V is a set of state variables, I is represents the initial state, O is the set of available operators/actions, and γ is the desired or goal state.

2.2 Single-Agent Action Plans

A single-agent action plan $\pi = (o_1, ...o_n)$ is a sequence of operators. o_1 is applicable from some initial state, and o_p is applicable after the application of o_{p-1} for $p > 1$.

2.3 Multi-Agent Planning Tasks

A multi-agent planning task is a 3-tuple $\Xi =< A, \Pi, T >$. Where $A = (A_1, ...A_k)$ is an ordered list of all agent names in the planning task. Π is the ordered list of agent planning tasks for each agent in A. $\Pi = (\Pi_1, ..., \Pi_n)$, and T is a scheduling function which determines which agents may act at any given time-point t.

2.4 Multi-Agent Action Plans

Multi-agent action plans are similar to single-agent action plans, but they also encode information about the agent which performed any given action [1].

A multi-agent action plan $\pi = ((A_x, \Pi_x[o]_1), ...(A_y, \Pi_y[o]_n))$. Where A_y is the label of the acting agent, and $\Pi_y[o]_n$ describes some applicable operator for agent y at time-point n.

For the purposes of readability, the rest of this paper will assume unique operator names and omit agent labels and specification of the operators available for that agent. Thus, multi-agent planning tasks will be treated as $\pi = (o_1, ...o_n)$, except where doing so would limit some required expressibility.

3 HP-Causality

This section briefly describes the notion of HP-causality and the three main definitions. @TODO must include 2 other definition and description of how inserting counterfactuals works.

3.1 The Modified definition of HP-Causality

Definition 3.1. $\vec{X} = \vec{x}$ is an actual cause of ϕ in the causal setting $(M, \vec{u})$ iff the following 3 conditions hold:

1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \psi$.
2. $AC2(a^m)$: There is a set $\vec{W}$ of variables in V and a setting $\vec{x'}$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w^*}$, then:
$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \vec{W} \leftarrow \vec{w^*}]\neg\phi$$
3. $\vec{X}$ is minimal. There is no strict subset of $\vec{X}$ that satisfies the previous 2 conditions.

4 HP-Causality Planning

The HP causality definitions above are restricted to the identification of variable causes, and as such, are not suitable for use in a planning domain where questions posed may also require the identification of a subset of actions or agents as causes.

For this reasons I propose changed definitions of the HP-causality models which produce comparable conclusions given intuitively identical, but structural different problems. Specifically, my definitions allow for the use of actions plans rather than structural equations as original used.

[1] I am aware that this information is trivially obtainable when agent actions have unique names, but I feel that it is more robust to also include explicit labelling information.

4.1 But-for Causality Planning

Definition 4.1. Given a multi-agent action plan π. The action subset $\vec{X} = \vec{x}$ is an but-for cause of ϕ in the causal setting $(M, \vec{u})$ iff the following 3 conditions hold:

1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \psi$.
2. $BF2$: There is a setting $\vec{x'}$ of the **alternative applicable actions** in $\vec{X}$ such that:
$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}]\neg\phi$$
3. $\vec{X}$ is minimal. There is no strict subset of $\vec{X}$ that satisfies the previous 2 conditions.

4.2 HP-Causality Modified Planning

Definition 4.2. Given a multi-agent action plan π. The action subset $\vec{X} = \vec{x}$ is an actual cause of ϕ in the causal setting $(M, \vec{u})$ iff the following 3 conditions hold:

1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \psi$.
2. $MC2(a^m)$: There is a set $\vec{W}$ of **actions** in π and a setting $\vec{x'}$ of the **alternative applicable actions** in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w^*}$, then:
$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \vec{W} \leftarrow \vec{w^*}]\neg\phi$$
3. $\vec{X}$ is minimal. There is no strict subset of $\vec{X}$ that satisfies the previous 2 conditions.

Agent Causes

Sometimes we may desire to know if a specific agent or subset of agents caused a particular variable assignment to occur. This allows for statements such as "Person X is solely responsible for event ϕ".

Definition 4.3. A subset of agents $\vec{A}$ is a cause of ϕ iff there exists some cause $\vec{X} = \vec{x}$ of ϕ such that every agent in $\vec{A}$ is a label of at least one action in $\vec{X}$, further, only agents mentioned in $\vec{A}$ are action labels in $\vec{X}$.

Proof. Complexity of determining if agent subset $\vec{A}$ is a cause of ϕ: (Assuming a known set of all action causes C).

1. Reduce all causes to the set of agent labels in their actions. Polynomial in the number of causes and the number of agents ($O(|C||A|)$).
2. Search through the set of cause labels for $\vec{A}$. Polynomial in the number of causes and the number of agents in the agent subset ($O(|C||\vec{A}|)$).

3. $O(|C||A|) \geq O(|C||\vec{A}|)$. Therefore $O(|C||A|)$ overall.

□

The above proof makes major assumptions about our ability to find C. An alternative approach is given here:

Proof. Complexity of determining if agent subset $\vec{A}$ is a cause of ϕ (Assuming a constant time complexity to determine if some variable set $\vec{X}$ is a cause of ϕ):

1. For an agent a to be a part of an agent cause, some action performed by a must be part of an action cause.
2. For a subset of agents to be a cause, all agents in that subset must have at least one action that is part of the same cause.
3. Thus, we must show that some combination of actions $\vec{X} = \vec{x} \in \pi$ containing at least one action from each agent is a cause of ϕ.
4. In this worst case this is related to finding (almost) all possible subsets of π (Π). Finding all possible subsets is exponential in the length of π, $O(2^{|\pi|})$.
5. Searching through these subsets to determine which contain one element from each agent is $O(|\vec{A}||\pi||\Pi|)$
6. Causation for all candidate sets in the restricted Π can be checked in $O(1)$ time from our assumptions.
7. $O(2^{|\pi|}) > O(|\vec{A}||\pi||\Pi|(1))$ Thus the time complexity is at least $O(2^{|\pi|})$.

□

Notes on the above definitions

- These definitions are strictly more inclusive than but-for cause checks. This is because when $\vec{W} = \{\}$, the $MC2(a^m)$ precisely models but-for causes (under minimality requirements for the but-for cause).
- The action plan obtained by replacing certain actions may no longer be valid. Much like in the normal HP definition of causality, where impossible variable assignments can be fixed during counter-factual reasoning.
- The most significant remaining concern relates to the treatment of exogenous events. Using a single exogenous event for all non-agent actions makes it very difficult to properly assign causality due to problems it causes with fixing variables. This will be detailed in the next section.

- One might wonder why my definitions do not refer to fixing variables at all, and only focus on actions. I have no strong argument for this decision except the observations that fixing some variable $X = x$ can be done in constant time by the inclusion of an action $s = (\top, X = x)$ in the action plan. This action can then be fixed. This addition can be done in $O(|X|)$ time.

4.3 Planning Causality relations

The following table shows which classes causes my definition of causality can identify. Columns are results ϕ. Rows are causes $\vec{X}$.

	Variable	**Action**
Variable	No[2]	No
Action	Yes	@TODO
Agent	Yes	@TODO

@TODO I would like to address the following questions:

1. What does it mean for an action to be the cause of an action? Is it sufficient to say that without that prior action the next action would not have occurred? What if only one action was available so $X = x' = x$?

2. How would extensions to a game theory approach affect causality? What if changing some action X would always result in a known change to some later action Y? Would we treat the two actions as a single action for the purposes of fixing and changing actions? How would this affect computational complexity?

5 Responsibility and Blameworthiness

5.1 HP Responsibility and Blameworthiness

5.2 Extending Responsibility and Blameworthiness to Planning

6 Examples

7 Conclusion