Causality in Multi-Agent Planning

Axel Ind

ALU-Freiburg

March 15, 2019

A Brief Overview

- What is causality?
- How did @TODOref formalize their definitions of causality?
- What is multi-agent planning?
- How can the work of @TODO ref be applied in a planning context?
- Some examples.

My Contributions

- Formalisation of a novel multi-agent AI planning framework for modelling causality.
- Introduction of counter-factual reasoning formalisms to the framework.
- Reinterpretations of the But-For, Original, Updated, and Modified causality definitions of @TODOref in the planning domain.

What is Causality?

- What does is really mean to say that A caused B?
- Commonly-used word, but our intuitions differ.
- Does is mean that without event B, effect A would never have happened?
- Broadly speaking, it means that we can identify the set of events that caused, or could have caused, some effect to occur, either in the real world or in some other possible world.

Alice's Homework

- "If Alice had not done her homework, she would have failed."
- Is it correct to say that not doing her homework caused Alice to fail?
- Do we implicitly assume that if she had done her homework she would have passed?
- Do we assume that not doing her homework is the soul cause of Alice failing?
- Are all the other pieces of homework that were never handed in also a cause of her failing?

Modelling Causality

- There are many intuitions about causal relations, and this is problematic.
- Evidently, we need to formalise our definitions of causality.
- But, which definitions should we use? Should they always match our intuition?
- In this presentation we will discuss four definitions of causality (as described by @TODOref).
- But-for Causality, HP's Original Causality, HP's Updated Causality, and HP's Modified causality.

Halpern's Causality

- @TODOref, like others before them recognized the inconsistencies in intuitive causality.
- Set about to systematically characterise a variety of causality intuitions.
- Purpose build framework for causality: Causal Settings.
- HP's definitions are unambiguous in this framework.

Counter-Factual Reasoning

- All definitions of causality ask one question: What if the world were other than it is?
- We need a way of expressing reasoning in a world where Suzy has done her homework.
- We need to be able to establish how changing some set of prior events (the cause) causes the counter-factual world to differ from the real world.

Causal Settings

- A causal setting (M, u) is description of the world limited to those variables that are relevant to cause and effect under consideration.
- M is the model, it describes the set of all variables and their interactions via structural equations. These variables are called *endogenous* variables.
- u is the set of all *exogenous* variables. These are variables whose values are determined by information outside of the scope of the model.

Structural Equations

- “If Alice had not done her homework, she would have failed.”
- One possible model:
  - $HW \leftarrow$ Alice does her homework.
  - $P \leftarrow$ Alice passes.
  - $u \leftarrow$ Exogenous variable.
  - $HW \leftarrow u$ We control if she does her homework from outside the model.
  - $P \leftarrow HW$ Doing her homework causes her to pass.

Causal Settings

- Now we can express the relationships in the model.
- But how do we evaluate the model?
- We define the causal model

$$(M, u) \models \phi$$

iff ϕ occurs after all structural equations in (M, u) have been exhaustively evaluated.

Causal Settings

@TODOdiagram of alice (including u values)

Causal Settings

- We can now describe the real world with casual settings.
- But we need counter-factual reasoning for causality.
- So we introduce the ability to modify variables:

  $$(M, u)[X \leftarrow x'] \models \phi$$

  iff ϕ holds after setting variables X to x' and evaluating the model.

Causal Settings

@TODOdiagram counter-factual of alice (including u values)

But-For Causality

- Now we are prepared to understand the first (and simplest) definition of causality.
- But-for Causality says: "$X = x$ is a cause of ϕ" if changing the value of X to x' means that ϕ no longer holds.

But-For Causality

Definition

$X = x$ is a but-for cause of ϕ in the causal setting (M, u) iff the following 3 conditions hold:

1. $(M, u) \models (X = x)$ and $(M, u) \models \phi$.
2. $AC2(bf)$: There is a setting x' of the variables in X such that

$$(M, u) \models [X \leftarrow x']\neg\phi$$

3. X is minimal, there is no strict subset of X' of X such that $X' = x'$ satisfies the above conditions.

Halpern’s Original Causality

- But-for causality catches only very simple intuitive causality.
- Halpern introduce his Original Causality to account for this deficiency.
- It asks two questions:
    1. Is there any setting of the variables in the model such that ϕ no longer holds.
    2. Is there any setting of variables in the model such that $X = x$ would be a but-for cause of ϕ.

Halpern's Original Causality

Definition

$X = x$ is an actual cause of ϕ in the causal setting (M, u) according to the original causality definition iff:

1. $(M, u) \models (X = x)$ and $(M, u) \models \phi$.
2. $AC2(a)$: There is a partition of V (the endogenous variables) into two disjoint subsets Z and W with $X' \subseteq Z'$ and a setting x' and w of the variables in X and W, respectively, such that

$$(M, u) \models [X \leftarrow x', W \leftarrow w]\neg\phi$$

3. $AC2(b^o)$: If $z^\star$ is such that $(M, u) \models Z = z^\star$, then for all subsets Z' of $Z - X$, we have

$$(M, u) \models [X \leftarrow x, W \leftarrow w, Z' \leftarrow z^\star]\phi$$

4. X is minimal, there is no strict subset of X' of X such that $X' = x'$ satisfies the above conditions.

Halpern's Updated Causality

- Eventually Halpern found cases in which even his Original definition of causality did not seem to capture human intuition (e.g. the voting scenario which we will discuss later.)
- He introduced his updated version of causality to ensure that the Original Causality definition held for every possible subset of W, W'.

Halpern's Updated Causality

Definition

$X = x$ is an actual cause of ϕ in the causal setting (M, u) according to the updated causality definition iff:

1. $(M, u) \models (X = x)$ and $(M, u) \models \phi$.
2. $AC2(a)$: There is a partition of V (the endogenous variables) into two disjoint subsets Z and W with $X' \subseteq Z'$ and a setting x' and w of the variables in X and W, respectively, such that

$$(M, u) \models [X \leftarrow x', W \leftarrow w]\neg\phi$$

3. $AC2(b^u)$: If $z^\star$ is such that $(M, u) \models Z = z^\star$, then for all subsets Z' of $Z - X$ and W' of W, we have

$$(M, u) \models [X \leftarrow x, W' \leftarrow w, Z' \leftarrow z^\star]\phi$$

4. X is minimal, there is no strict subset of X' of X such that $X' = x'$ satisfies the above conditions.

Halpern's Modified Causality

- Finally, we come to Halpern's Modified Causality.
- This definition is Halpern's preferred causality, and has the added advantage of being far simpler that the Original and Updated definitions.
- This definition is a direct extension of the But-For cause definition of causality.
- It additionally allows the $X = x$ is a cause if fixing any subset of other variable assignments to their original values would cause $X = x$ to be a But-For cause of ϕ.

Halpern's Modified Causality

Definition

$X = x$ is an actual cause of ϕ in the causal setting (M, u) iff the following 3 conditions hold:

1. $(M, u) \models (X = x)$ and $(M, u) \models \phi$.
2. $AC2(a^m)$: There is a set $\vec{W}$ of variables in V and a setting x' of the variables in X such that if $(M, u) \models W = w^*$, then:

$$(M, u) \models [X \leftarrow x', W \leftarrow w^*]\neg\phi$$

3. X is minimal. There is no strict subset of X that satisfies the previous 2 conditions.

Causality in Planning

- We have now covered the principles of modelling causality.
- Now we approach the idea of causality in a Multi-agent AI Planning context.

Why is Planning Causality Different?

- The domain differs in that the causes under consideration are no longer variables, but actions. (Although the effects ϕ remain variables.)
- Planning also exerts a constraint that HP Causality does not. Actions must be performed *sequentially*.
- There is no independence of actions. Counterfactual reasoning with action a_i may later render action a_{i+k} inapplicable.

What Extra Value does Causality Planning Offer?

- AI Planning is an extensively studied domain,
- We do not need to interpret our models in the Causal Setting to determine causality. (Such domain changing could well invalidate the conclusions anyway.)
- We are no longer constrained to independent events.

Single-Agent Planning Tasks

@TODOdiagram

Multi-agent Planning Tasks

@TODOdiagram

Exogenous Actions

- Exogenous actions are those actions which are not performed by any agent.
- Their occurrence can be timed, after each agent action, or a combination of these factors. (We use timed events.)
- Unlike @TODOref we allow conflicting exogenous actions. (Without timed events this leads to multiple possible final states.)

Plans

- A plan π is a sequence of actions.
- Each action at position i, a_i must be applicable after the application of a_{i-1}, $i \geq 1$, and π_0 must be applicable in the initial state.
- For use in counterfactual reasoning, we define *action slots*. An action slot $q(\pi, k)$ is mapped to the list of all applicable actions at position k in the plan. Intuitively an action slot functions like a variable name for the domain of possible actions at a specific time-point in π.

Causal Plan Settings

- As with HP Causality we require a way to describe the world prior to the execution of the model, and the world that results afterwards.
- For that reason we define the Casual Plan Setting (π, Π).
- The causal Plan Setting (CPS) contains all relevant information for causality modelling.
- We write

  $$(\pi, \Pi) \models (\Diamond \phi)$$

  iff execution of the plan results in the variable assignments ϕ in the final state.

Causal Plan Settings

- As with causal settings counter-factual reasoning is possible in plan settings too.
- We say

$$(\Pi, \pi)[\pi' \leftarrow o'] \models \phi$$

  iff counterfactually using actions o' in the CPS would result in ϕ.
- This definition might seem identical to that we saw in causal plans, but there are key differences.

CPS New Operations

- Counterfactual reasoning in causal planning settings is more complex than in causal settings.
- Unlike is causal setting we have no guarantee that plan will remain applicable after changing some action in π.
- Here we follow @TODO ref and treat any counter-factual assignment as applicable, provided the original plan was applicable.

CPS New Operations

- Action slots were introduced to handle counter-factual reasoning in a paradigm where the order of action executions matters.
- A subplan π', $\pi' \subset \pi$ is a subset of the action slots in π.
- Thus a possible subplan of $\pi = (a_1, a_2, a_3)$ is $\pi' = \{q(\pi, 1), q(\pi, 3)\}$ which tells us all actions that may have occurred in the first and last position of π.
- $\pi - \pi'$ returns a list of all action slots that occur in π and not in π'. @TODOdiagram
- Remember that a plan and a subplan are not interchangeable because a plan associated a single action with each action slot.

But-For Causality Planning

- Now we can finally discuss causality planning.
- The intuition of But-For causality does not change, and we now simply try to determine “Would changing π' from $\pi' = o$ to $\pi' \leftarrow o'$ change the final value of some variables ϕ?”

But-For Causality Planning

Definition

Given a multi-agent action plan π with a final state s_n, some action slots in π, $\pi' \subseteq \pi$, $\pi' \leftarrow o$ are a but-for cause of some final variable assignment $s_n \models \phi$ iff the following 3 conditions hold:

1. $o \subseteq \pi$, $\pi' \models o$ and $s_n \models \phi$.
2. $BF2$: There is a setting o' of the non-ϵ action slots in π' such that:
$$(\pi, \Xi) \models [\pi' \leftarrow o'](\neg \Diamond \phi)$$
3. π' is minimal. There is no strict subset of π' that satisfies the previous 2 conditions.

Original Causality Planning

- Capturing the intuition of the original definition of HP causality is more difficult.
- We now restrict our counter-factual reasoning to endogenous actions.

Original Causality Planning

Definition

Given a multi-agent planning task Ξ and action plan π with a final state s_n, some actions slots π', $\pi' \subseteq \pi$, $\pi' \leftarrow o$ are a cause of some final variable assignment $s_n \models \phi$ according to the Original planning definition iff:

1. $o \subseteq \pi$, $\pi' \models o$ and $s_n \models \phi$.
2. $BF2(a)$: There is a partition of F (the *endogenous* actions) into two disjoint subplans $Z \subseteq \pi$ and $W \subseteq (\pi - Z)$ with $\pi' \subseteq Z$ and a setting o' and w' of the actions in π' and W respectively such that

   $$(\pi, \Xi) \models [\pi' \leftarrow o', W \leftarrow w'](\neg\Diamond\phi)$$

3. $BF2(b^o)$: If $z^\star$ is such that $(\pi, \Xi) \models Z = z^\star$, then for all subplans $Z' \subseteq (Z - \pi')$ we have

   $$(\pi, \Xi) \models [\pi \leftarrow o, W \leftarrow w', Z' \leftarrow z^\star](\Diamond\phi)$$

4. π' is minimal. There is no strict subset of π' that satisfies the previous 3 conditions.

Updated Causality Planning

- We will not have time to discuss Updated Causality planning in detail, but it is sufficent to note that Original planning failed in the same cases as HP's Original Casuality and Updated Planning addresses the exact same circumstances as HP's Updated Causality.

Updated Causality Planning

Definition

Given a multi-agent planning task Ξ and action plan π with a final state s_n, some actions slots π', $\pi' \subseteq \pi$, $\pi' \leftarrow o$ are a cause of some final variable assignment $s_n \models \phi$ according to the Updated planning definition iff:

1. $o \subseteq \pi$, $\pi' \models o$ and $s_n \models \phi$.
2. $BF2(a)$: There is partition of F (the endogenous actions) into two disjoint subplans $Z \subseteq \pi$ and $W \subseteq (\pi - Z)$ with $\pi' \subseteq Z$ and a setting o' and w' of the actions in π and W respectively such that

$$(\pi, \Xi) \models [\pi' \leftarrow o', W \leftarrow w'](\neg\Diamond\phi)$$

3. $BF2(b^u)$: If $z^\star$ is such that $(\pi, \Xi) \models Z = z^\star$, then *for all* subplans $W' \subseteq W$ and $Z' \subseteq (Z - \pi')$ we have

$$(\pi, \Xi) \models [\pi \leftarrow o, W' \leftarrow w', Z \leftarrow z^\star](\Diamond\phi)$$

4. π' is minimal. There is no strict subset of π' that satisfies the previous 3 conditions.

Modified Causality Planning

- Finally, we change the modified version of causality to the planning domain.
-

Modified Causality Planning

Definition

Given a multi-agent action plan π with a final state s_n, some action slots π',$\pi' \subseteq \pi$, $\pi' \leftarrow o$ are a cause of some final variable assignment $s_n \models \phi$ according to the modified planning definition iff the following 3 conditions hold:

1. $o \subseteq \pi$, $\pi' \models o$ and $s_n \models \phi$.
2. $BF2$: There is a setting o' of the applicable actions in π', and a setting of $W \subseteq (\pi - \pi')$ such that:

$$(\pi, \Xi) \models [\pi' \leftarrow o', W \leftarrow w^{\star}](\neg\Diamond\phi)$$

3. π' is minimal. There is no strict subset of π' that satisfies the previous 2 conditions.

(Where $W \leftarrow w^{\star}$ denotes fixing all non-ϵ action slots in W to their original actions.)

Examples

Example 1: Forest Fire

Example 2: Rock Throwing

Example 3: Voting

Conclusions

Future Work

Special Thanks