Causality in Multi-Agent Planning

Axel Ind

ALU-Freiburg

March 14, 2019

Alice

- This is Alice.
- Alice is studying.
- But she's not a very good student.

@TODOimage

Alice

- Alice did not hand in any homework.
- Alice failed her class.
- Why did Alice fail?

@TODOimage

Bob

- This is Bob.
- Bob likes fires.
- ... A little too much.
- Bob is an arsonist.
- And he has found a box of matches.

@TODOimage

Bob

- Bob is in a forest.
- He drops a lit match.
- At the same moment, lightning strikes.
- The forest burns down.
- Why did the forest burn down?

@TODOimage

Alice and Bob

- Alice and Bob are friends.
- They find an empty house.
- They decide to throw stones at the windows.

@TODOimage

Alice

- They each throw a stone.
- The window shatters.
- Who should we blame?

@TODOimage

A Brief Overview

- What is causality?
- Halpern's Causality.
- Counterfactaul reasoning in planning.
- Planning Causality.

My Contributions

- Multi-agent framework for causality.
- Four types of causality in planning.

@TODOdiagramlinediagram maybe venn diagram?

What is Causality?

- A caused B.
- Our intuitions differ.
- @TODO ref formalised their own definitions.

Alice's Homework

- "If Alice had not done her homework, she would have failed."
- Possible cause: not doing her homework.
- What if she *had* done it?

@TODOdiagram

Modelling Causality

- Definitions should match our intuition.
- Four Definitions of Causality.
- But-for Causality, HP's Original Causality, HP's Updated Causality, and HP's Modified causality.

@TODOdiagram of all 4 types

Halpern's Causality

Halpern's Causality

- Introduced the causal setting.
- Allows counter-factual reasoning.
- Unambiguous causality in this framework.

Causal Settings

- Causal setting (M, u).
- Model M.
- Exogenous variables u.
- Describes the initial and final state of the world.

@TODOexampleimage

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.

@TODOimage replace

Causal Settings

@TODOdiagram of alice (including u values) and counterfactuals

But-For Causality

- "$X = x$ 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.

Example 1: Alice's Homework

@TODOdiagram

Example 2: Bob's Fire

@TODOdiagram

Example 3: Alice and Bob's Fire

@TODOdiagram

Halpern's Modified Causality

- But-for causality is limited.
- Extends but-for causality.

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.

Example 1: Alice's Homework

@TODOdiagram

Example 2: Bob's Fire

@TODOdiagram

Example 3: Alice and Bob's Fire

@TODOdiagram

Test section one

Causality in Planning

- Causal settings require independent structural equations.
- Planning is more expressive.
- Idea: extend HP's Causality for AI Planning.

Why is Planning Causality Different?

- Actions and variables are distinct.
- Exogenous actions.
- Sequential ordering of actions.
- Identify agent causes.

Single-Agent Planning Tasks

@TODOdiagram

Multi-agent Planning Tasks

@TODOdiagram

Exogenous Actions

- Not performed by any agent.
- Can be timed, or we get non-determinism.
- Unlike @TODOref we allow conflicting exogenous actions.

Plans

- A plan π is a sequence of actions.
- We define *action slots*.

Definition

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.

Plans

@TODOdiagram action slot example

Causal Plan Settings

- We define the Casual Plan Setting (π, Π).
- Similar intuition to causal settings.

$$(\pi, \Pi) \models (\text{finally} \phi)$$

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

$$(\Pi, \pi)[\pi' \leftarrow o'] \models (\text{finally} \phi)$$

iff counterfactually using actions o' in the CPS would result in ϕ.

CPS New Operations

- A subplan π', $\pi' \subset \pi$ is a subset of the action slots in π.
- A plan and a subplan are not interchangeable.

@TODOdiagram

But-For Causality Planning

- "Would changing π' from $\pi' = o$ to $\pi' = o'$ change the final value of some variables ϕ?"

But-For Causality Planning

Definition

Given a multi-agent action plan Ξ and a 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. *BF*2: There is a setting o' of the non-ϵ action slots in π' such that:

$$(\pi, \Xi) \models [\pi' \leftarrow o'](\neg \text{finally} \phi)$$

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

Modified Causality Planning

Definition

Given a multi-agent action plan Ξ and 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.)

Example 1: Alice's Homework

@TODOdiagram

Example 2: Bob's Fire

@TODOdiagram

Example 3: Alice and Bob's Fire

@TODOdiagram

Conclusions

Future Work

Special Thanks

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.

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