

Causality in Multi-Agent Planning

Axel Ind

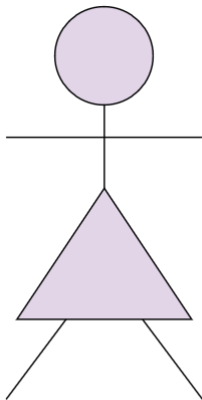
ALU-Freiburg

March 15, 2019

Halpern's Causality

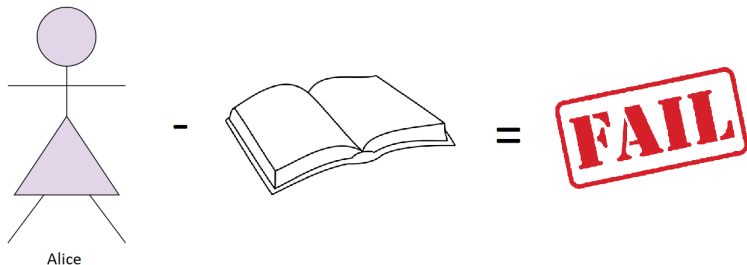
Alice

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



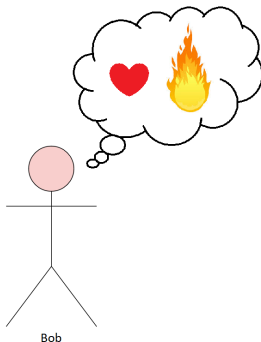
Alice

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

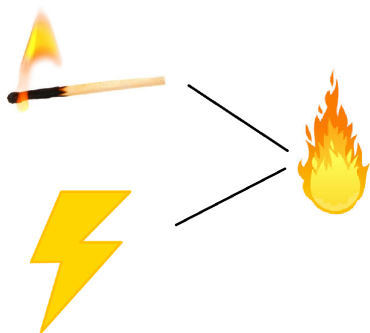


Bob

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

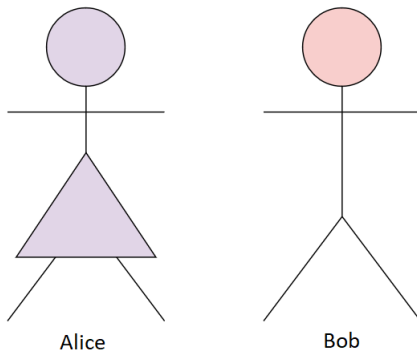


- Bob is in a forest.
- He drops a lit match.
- At the same moment, lightning strikes.
- The forest burns down.
- Why did the forest burn down?



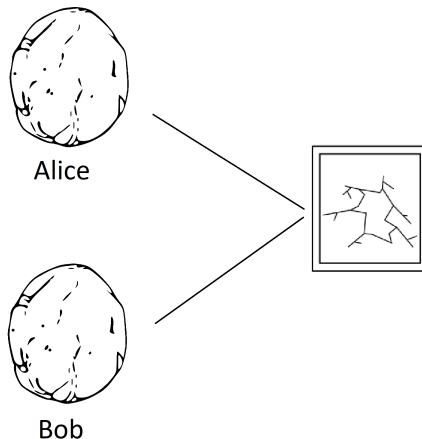
Alice and Bob

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



Alice

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

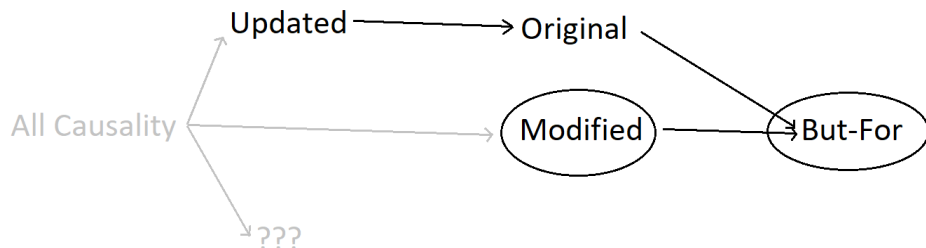


A Brief Overview

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

My Contributions

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

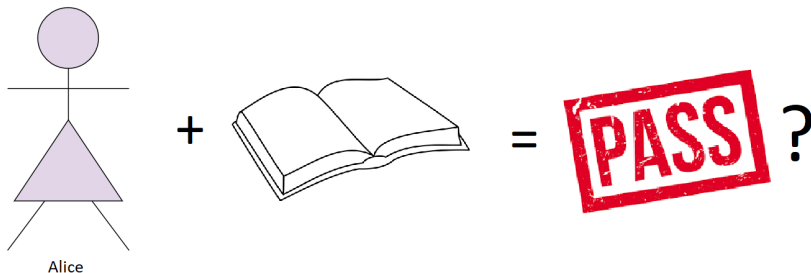


What is Causality?

- 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?



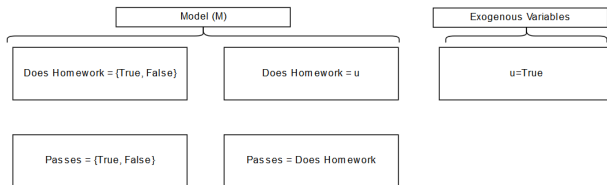
Halpern's Causality

Halpern's Causality

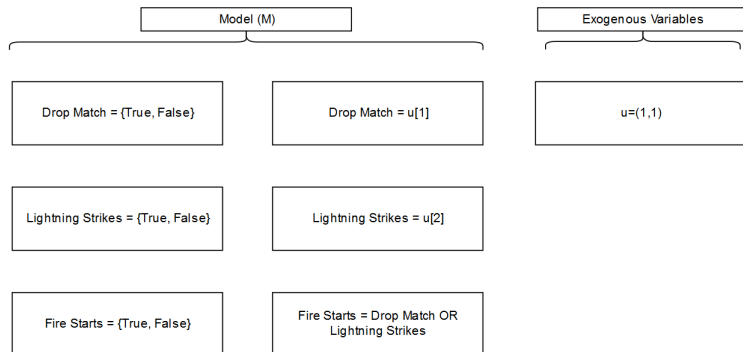
- The causal setting.
- Allows counter-factual reasoning.
- Unambiguous causality in this framework.

Causal Settings: Alice's Homework

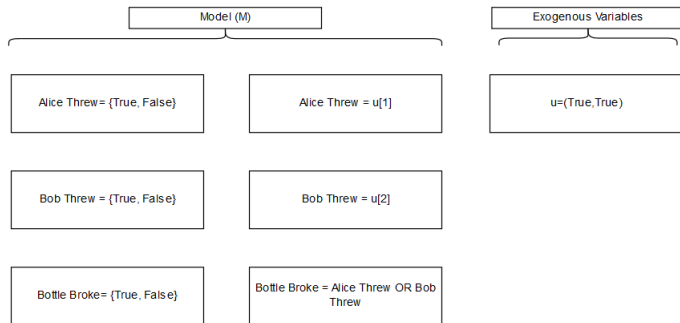
- Causal setting (M, u) .
- Sufficient to the initial and final state of the world.



Causal Settings: Bob's Fire

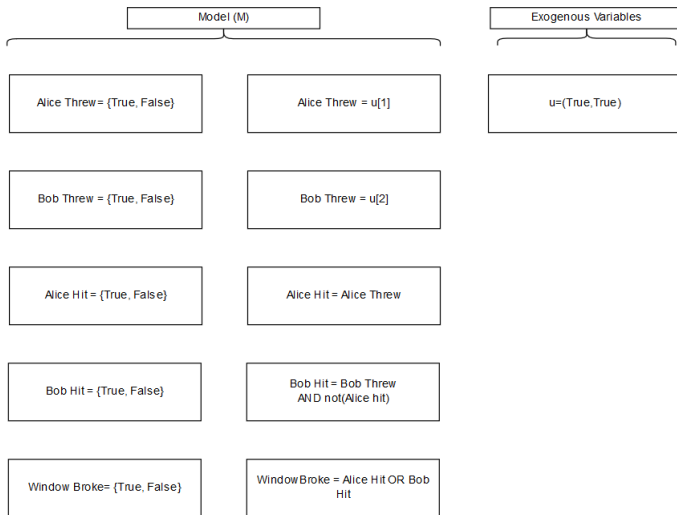


Causal Settings: Alice and Bob



- Identical to the Forest Fire example.
- How can we change it?

Causal Settings: Alice and Bob



- “ $X = x$ is a cause of ϕ ” if changing the value of X to x' means that ϕ no longer holds.

Definition

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

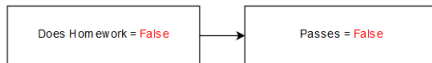
- ① $(M, u) \models (X = x)$ and $(M, u) \models \phi$.
- ② $AC2(bf)$: There is a setting x' of the variables in X such that

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

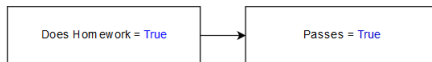
- ③ X is minimal, there is no strict subset X' of X such that $X' = x'$ satisfies the above conditions.

Example 1: Alice's Homework

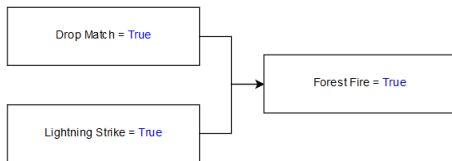
- In the actual world she didn't do her homework and failed.



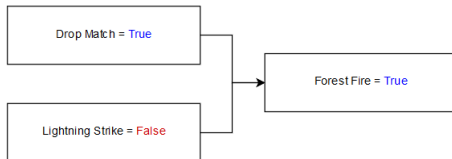
- In our model, counter-factually examining the world in which she did her homework results in her passing.



Example 2: Bob's Fire

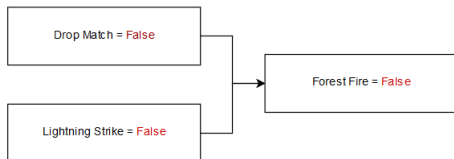


- Counter-factually stating that Bob did not drop the match still results in the forest fire.



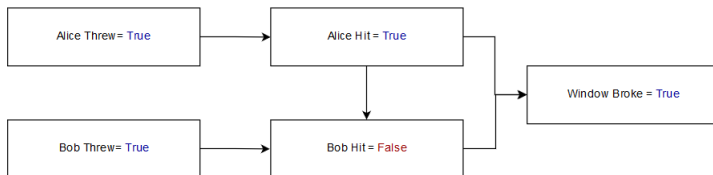
Example 2: Bob's Fire

- If neither the match was dropped nor the lightning struck, there would have been no forest fire.

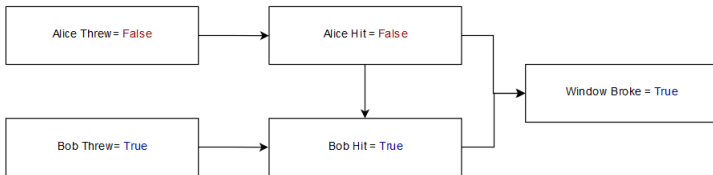


- $X = \{DropMatch = True\}$ is not a cause of *forestfire*, but $X = \{DropMatch = True, LightningStrike = True\}$ is a cause.

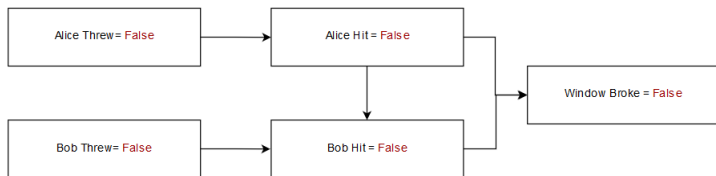
Example 3: Alice and Bob's Window



- With $x' = \{AliceThrew = False\}$, Bob's stone now breaks the window.



Example 3: Alice and Bob's Window



- $X = \{AliceThrew = True\}$ is not a cause of $BottleBroke = True$.
- $X = \{AliceThrew = True, BobThrew = True\}$ is a cause of $BottleBroke = True$.

Halpern's Modified Causality

- Intuitively we might feel that Alice is a cause, but Bob is not.
- But-for causality is limited.
- Extends but-for causality.

Definition

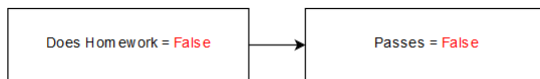
$X = x$ is an actual cause of ϕ in the causal setting (M, u) iff the following 3 conditions hold:

- 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:

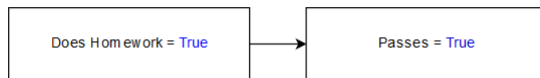
$$(M, u) \models [X \leftarrow x', W \leftarrow w^*] \neg \phi$$

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

Example 1: Alice's Homework

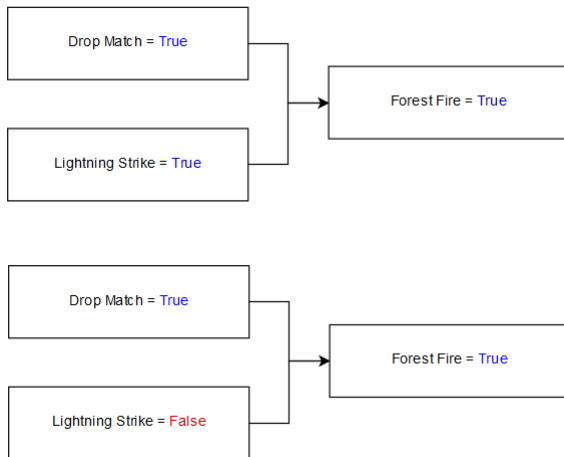


- A But-For cause is a cause in the Modified definition with $W = \{\}$.



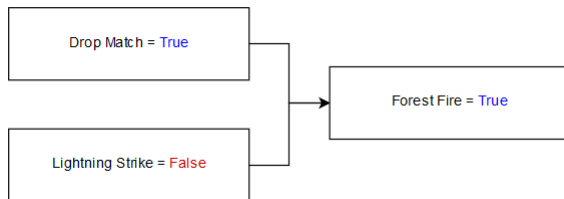
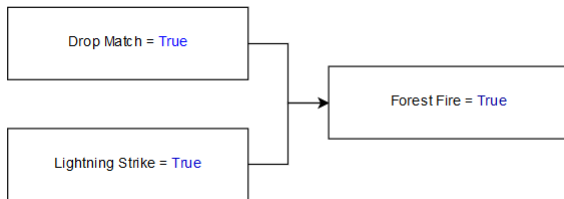
- $X = \{DoesHomework = True\}$ is a cause in Modified Definition.

Example 2: Bob's Fire

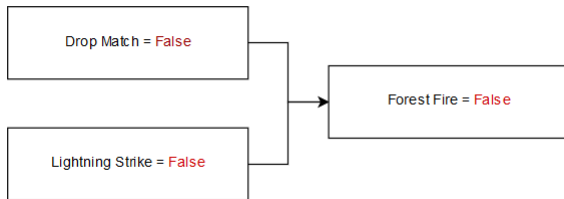


- No matter what other variables we find as W , we can't cause *ForestFire* = *False*.

Example 2: Bob's Fire

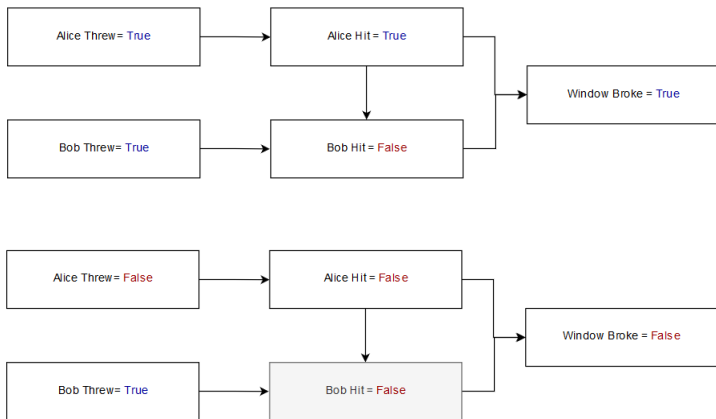


Example 2: Bob's Fire



- $X = \{DropMatch = True, LightningStrike = True\}$, with $W = \{\}$ is a cause of $ForestFire = True$

Example 3: Alice and Bob's Window



- By fixing $W = \{BobHit = True\}$, $X = \{AliceThrew = True\}$ is cause of $WindowBroke = True$.

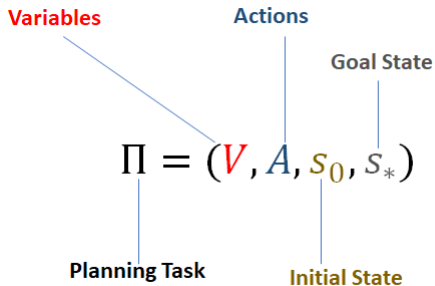
Causality in Planning

- 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



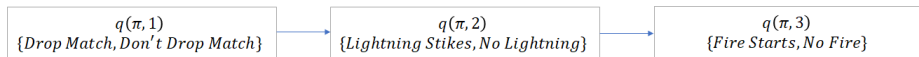
- Not performed by any agent.
- Can be timed, or we get non-determinism.
- Unlike @TODOref we allow conflicting exogenous actions.

- 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 .

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



- Valid π' values: $\{q(\pi', 1), q(\pi', 2), q(\pi', 3)\}$, $\{q(\pi', 1), q(\pi', 2)\}$, $\{q(\pi', 1), q(\pi', 3)\}$, $\{q(\pi', 2), q(\pi', 3)\}$, $\{q(\pi', 2), q(\pi', 3)\}$, $\{(\pi', 1)\}$, $\{(\pi', 2)\}$, $\{(\pi', 3)\}$

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 .

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

Definition

Given a planning task Π 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 *BF2*: There is a setting o' of the non- ϵ action slots in π' such that:

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

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

Example 1: Alice's Homework



- What other actions could Alice have taken taken?
- $\pi' = q(\pi, 1)$, $o' = (Don't Do Homework)$



Example 2: Bob's Fire



- $\pi' = q(\pi, 1), o' = \{DropMatch\}$



- $\pi' = \{q(\pi, 1), q(\pi, 2)\}, o' = \{DropMatch, LightningStrikes\}$



Example 3: Alice and Bob's Window



- $\pi' = \{q(1, \pi)\}, o' = \{NotAliceThrows\}$



- $\pi' = \{q(1, \pi) q(2, \pi)\}, o' = \{NotAliceThrows, NotBobThrows\}$



Definition

Given a planning task Π and plan π with a final state s_n , some action slots $\pi', \pi' \subseteq \pi$, $\pi' \leftarrow o$ 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:

$$(\pi, \Pi) \models [\pi' \leftarrow o', W \leftarrow w^*](\neg \text{finally} \phi)$$

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

Example 1: Alice's Homework



- $\pi' = q(\pi, 1)$, $o' = (Don't Do Homework)$, $W = \{\}$



Example 2: Bob's Fire



- $\pi' = \{q(\pi, 1)\}, o' = \{Don't Drop Match\}, W = \{\}$



- $\pi' = \{q(1, \pi) q(2, \pi)\},$
 $o' = \{Not Alice Throws, Not Bob Throws\}, W = \{\}$



Example 3: Alice and Bob's Window



- $\{q(\pi, 1)\}$, $o' = \{NotAliceThrows\}$, $W = \{q(\pi, 3)\}$,
 $w^* = \{NotHitsBob\}$



- @TODO ref's causality.
- Extending Causality.
- 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^* is such that $(M, u) \models Z = z^*$, then for all subsets Z' of $Z - X$, we have

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

- 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^* is such that $(M, u) \models Z = z^*$, then for all subsets Z' of $Z - X$ and W' of W , we have

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

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

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

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^* is such that $(\pi, \Xi) \models Z = z^*$, then for all subplans $Z' \subseteq (Z - \pi')$ we have

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

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

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

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^* is such that $(\pi, \Xi) \models Z = z^*$, then *for all* subplans $W' \subseteq W$ and $Z' \subseteq (Z - \pi')$ we have

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

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

Questions