From verification to causality-based explications

Christel Baier

TU Dresden

Joint work with:

Clemens Dubslaff
Florian Funke
Stefan Kiefer
Simon Jantsch
Rupak Majumdar
Corto Mascle
Jakob Piribauer
Robin Ziemek

From verification to explications

Classical verification task:

given: a system model $\mathcal{M}$ and a specification ϕ

question: does $\mathcal{M}$ satisfy ϕ?

From verification to explications

Classical verification task:

given:	a system model $\mathcal{M}$ and a specification ϕ
question:	does $\mathcal{M}$ satisfy ϕ?
answer:	yes or no

From verification to explications

Classical verification task:

given: a system model $\mathcal{M}$ and a specification ϕ

question: does $\mathcal{M}$ satisfy ϕ?

answer: yes or $\boxed{\text{no}}$

counterexample

From verification to explications

Classical verification task:

given: a system model $\mathcal{M}$ and a specification ϕ

question: does $\mathcal{M}$ satisfy ϕ?

answer: $\boxed{\text{yes}}$ or no

mathematical proof or certificate

counterexample

From verification to explications

Classical verification task:

given:	a system model $\mathcal{M}$ and a specification ϕ
question:	does $\mathcal{M}$ satisfy ϕ?
answer:	yes or no

mathematical proof or certificate → yes

counterexample → no

Explication task (in the verification context):

... should provide deeper insights **why** the specification holds or not

From verification to explications

Classical verification task:

given: a system model $\mathcal{M}$ and a specification ϕ

question: does $\mathcal{M}$ satisfy ϕ?

answer: yes or no

mathematical proof or certificate → yes

counterexample → no

Explication task (in the verification context):

- what causes the specification to hold for the full model ?
- who is responsible for a requirement violation ? and to which degree?
- if a bad behavior occurs, what has caused the violation of the specification ?

From verification to explications

Classical verification task:

given: a system model $\mathcal{M}$ and a specification ϕ

question: does $\mathcal{M}$ satisfy ϕ?

answer: yes or no

mathematical proof or certificate

counterexample

Explication task (in the verification context):

- what causes the specification to hold or not?
- who is responsible for a requirem [illegible] degree?
- if a bad behavior occurs, what has caused the violation of the specification ?

"causality meets verification"

Causality

long-standing discussion in philosophy

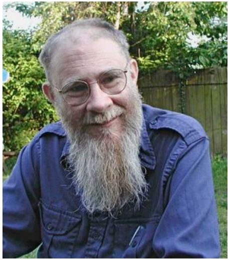
David Hume

(philosopher, 1711-1776)

painting from
Allan Ramsay

David K. Lewis

(philosopher, 1941-2001)



By Source, Fair use,
https://en.wikipedia.org/w/index.php?curid=58724625

and many more ...

Causality

long-standing discussion in philosophy, but also AI

Joseph Halpern

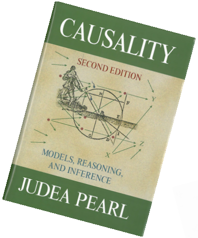
Gödel Prize 1997
Dijkstra Prize 2009

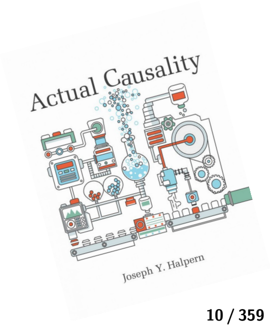
Joe Halpern at EPFL in June 2008

Judea Pearl

Turing Award
Winner 2011

taken from Judea Pearl's homepage
UCLA Cognitive Systems Laboratory





Various forms of causality

Various forms of causality

- actual/specific vs general/type causes

  actual cause is a factual event C that causes the effect E

  general cause: e.g. “sweets cause obesity”

Various forms of causality

- actual/specific vs general/type causes

  actual cause is a factual event C that causes the effect E

  general cause: e.g. "sweets cause obesity"

- backward vs forward causality-based reasoning

  backward: what has caused an observed effect E in a given event sequence?

  forward: what can cause an event E in a given world model?

Various forms of causality

- actual/specific vs general/type causes

  actual cause is a factual event C that causes the effect E

  general cause: e.g. "sweets cause obesity"

- backward vs forward causality-based reasoning

  backward: what has caused an observed effect E in a given event sequence?

  forward: what can cause an event E in a given world model?

- counterfactual vs necessary vs sufficient cause-effect relations

  counterfactual: if C would not have happened, then E would not have occured

  necessary: if E occurs then C must have happened before

  sufficient: if C happens then always E will occur somewhen later

Various forms of causality

- actual/specific vs general/type causes

  actual cause is a factual event C that causes the effect E

  general cause: e.g. "sweets cause obesity"

- backward vs forward causality-based reasoning

  backward: what has caused an observed effect E in a given event sequence?

  forward: what can cause an event E in a given world model?

- counterfactual vs necessary vs sufficient cause-effect relations

  counterfactual: if C would not have happened, then E would not have occured

  necessary: if E occurs then C must have happened before

  sufficient: if C happens then always E will occur somewhen later

- deterministic vs probabilistic causes, and many more ...

Causality in the verification context

Causality in the verification context

- program slicing [Weiser'79]

which program statements affect the values of variables
at a certain program location?

Causality in the verification context

- program slicing [Weiser'79]
- causality-based explanations of counterexamples
  - ⋆ counterfactual reasoning with distance metrics [Groce et al'06]
  - ⋆ identification of "critical state-variable pairs" in cex [Beer et al'09]
  - ⋆ event order logic for causal dependencies in cex [Leitner-Fischer/Leue'13]

    ⋮

Causality in the verification context

- program slicing [Weiser'79]
- causality-based explanations of counterexamples
  - ⋆ counterfactual reasoning with distance metrics [Groce et al'06]
  - ⋆ identification of "critical state-variable pairs" in cex [Beer et al'09]
  - ⋆ event order logic for causal dependencies in cex [Leitner-Fischer/Leue'13]
- coverage and vacuity [Chockler et al'01, Beer et al'01, Kupferman/Vardi'03]
  study mutations of system models or specifications

Causality in the verification context

- program slicing [Weiser'79]
- causality-based explanations of counterexamples
  - ⋆ counterfactual reasoning with distance metrics [Groce et al'06]
  - ⋆ identification of “critical state-variable pairs” in cex [Beer et al'09]
  - ⋆ event order logic for causal dependencies in cex [Leitner-Fischer/Leue'13]
- coverage and vacuity [Chockler et al'01, Beer et al'01, Kupferman/Vardi'03]
  study mutations of system models or specifications
- causality and responsibility in operational models
  - ⋆ cause-effect relations [Cho./Hal./Kup.'08, B./Fun./Maj.'21, B./Fun./Pir./Zie.'22]
  - ⋆ quantitative measures for the relevance of states [Chockler/Halpern/Kupf.'08, B./Funke/Maj.'21, Mascle et al'21]

Causality in the verification context

- program slicing [Weiser'79]
- causality-based explanations of counterexamples
  - ⋆ counterfactual reasoning with distance metrics [Groce et al'06]
  - ⋆ identification of “critical state-variable pairs” in cex [Beer et al'09]
  - ⋆ event order logic for causal dependencies in cex [Leitner-Fischer/Leue'13]
- coverage and vacuity [Chockler et al'01, Beer et al'01, Kupferman/Vardi'03]
  study mutations of system models or specifications
- causality and responsibility in operational models
  - ⋆ cause-effect relations [Cho./Hal./Kup.'08, B./Fun./Maj.'21, B./Fun./Pir./Zie.'22]
  - ⋆ quantitative measures for the relevance of states
    [Chockler/Halpern/Kupf.'08, B./Funke/Maj.'21, Mascle et al'21]
- causality-based verification [Kupriyanov/Finkbeiner'13]
  proof rules for stepwise cause-effect reasoning

Causality in the verification context

- program slicing [Weiser'79]
- causality-based explanations of counterexamples
  - ⋆ counterfactual reasoning with distance metrics [Groce et al'06]
  - ⋆ identification of "critical state-variable pairs" in cex [Beer et al'09]
  - ⋆ event order logic for causal dependencies in cex [Leitner-Fischer/Leue'13]
- coverage and vacuity [Chockler et al'01, Beer et al'01, Kupferman/Vardi'03]
  study mutations of system models or specifications
- causality and responsibility in operational models
  - ⋆ cause-effect relations [Cho./Hal./Kup.'08, B./Fun./Maj.'21, B./Fun./Pir./Zie.'22]
  - ⋆ quantitative measures for the relevance of states
    [Chockler/Halpern/Kupf.'08, B./Funke/Maj.'21, Mascle et al'21]
- causality-based verification [Kupriyanov/Finkbeiner'13]
  proof rules for stepwise cause-effect reasoning

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
- Probabilistic causality in Markovian models
- Conclusions

Cause-effect relations in TS

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Define forward notions of causality:

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Define forward notions of causality:

- necessary cause

  "if the effect occurs then the cause must have happened before"

- sufficient cause

  "if the cause happens then the effect will occur somewhen later"

- counterfactual cause

  "set of states with minimal number of modifications to avoid the effect"

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Define forward notions of causality:

- necessary cause

  "if the effect occurs then the cause must have happened before"

- sufficient cause

  "if the cause happens then the effect will occur somewhen later"

- counterfactual cause

  "set of states with minimal number of modifications to avoid the effect"

... many possible formalizations ...

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Define forward notions of causality:

- necessary cause

  "if the effect occurs then the cause must have happened before"

- sufficient cause

  "if the cause happens then the effect will occur somewhen later"

- counterfactual cause

  "set of states with minimal number of modifications to avoid the effect"

Here: characterization of necessary/sufficient causes using CTL*

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. C is called a

- necessary cause for E if $\mathcal{M} \models \forall \neg \big((\neg C) \,\mathsf{U}\, E \big)$

  "if the effect occurs then the cause must have happened before"

- sufficient cause ...

  "if the cause happens then the effect will occur somewhen later"

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. C is called a

- necessary cause for E if $\mathcal{M} \models \forall\neg\big((\neg C)\,\mathsf{U}\,E\big)$

  "if the effect occurs then the cause must have happened before"

- sufficient cause for E if $\mathcal{M} \models \forall\Box\big(C \rightarrow \bigcirc\Diamond E\big)$

  "if the cause happens then the effect will occur somewhen later"

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. C is called a

- necessary cause for E if $\mathcal{M} \models \forall\neg((\neg C)\,\mathsf{U}\,E)$

  "if the effect occurs then the cause must have happened before"

- sufficient cause for E if $\mathcal{M} \models \forall\Box(C \rightarrow \bigcirc\Diamond E)$

  "if the cause happens then the effect will occur somewhen later"

Monotonicity:

$$C \text{ is necessary and } C \subseteq D \implies D \text{ is necessary}$$

$$C \text{ is sufficient and } C \supseteq D \implies D \text{ is sufficient}$$

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. C is called a

- necessary cause for E if $\mathcal{M} \models \forall\neg((\neg C)\,\mathsf{U}\,E)$

  "if the effect occurs then the cause must have happened before"

- sufficient cause for E if $\mathcal{M} \models \forall\Box(C \rightarrow \bigcirc\Diamond E)$

  "if the cause happens then the effect will occur somewhen later"

Transitivity (up to disjointness):

C necessary for D & D necessary for E $\Longrightarrow$ C necessary for E

C sufficient for D & D sufficient for E $\Longrightarrow$ C sufficient for E

Cause-effect relations in TS

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states.

Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. C is called a

- necessary cause for E if $\mathcal{M} \models \forall \neg((\neg C) \, \mathsf{U} \, E)$

  "if the effect occurs then the cause must have happened before"

- sufficient cause for E if $\mathcal{M} \models \forall \Box (C \rightarrow \bigcirc \Diamond E)$

  "if the cause happens then the effect will occur somewhen later"

If all E-states are terminal then:

$$C \text{ is necessary} \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$$

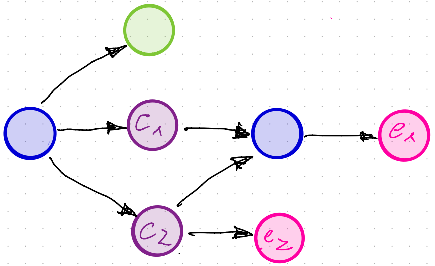
$$C \text{ is sufficient} \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$$

Example: necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$

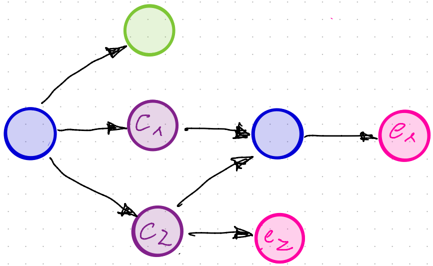


effect set $E = \{e_1, e_2\}$

Example: necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

> C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$
>
> C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$



effect set $E = \{e_1, e_2\}$

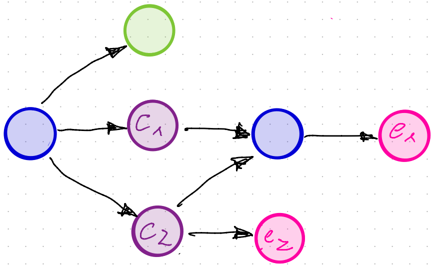
$\{c_1, c_2\}$ necessary and sufficient

Example: necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \to \Diamond C)$

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \to \Diamond E)$



effect set $E = \{e_1, e_2\}$

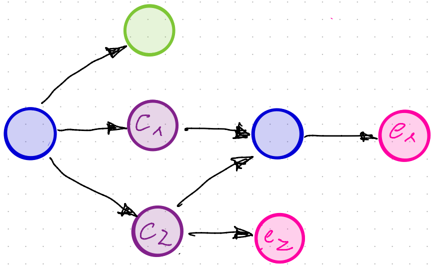
$\{c_1, c_2\}$ necessary and sufficient

$\{c_1\}$ sufficient, not necessary

Example: necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E	iff	$\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$
C is a sufficient cause for E	iff	$\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$



effect set $E = \{e_1, e_2\}$

$\{c_1, c_2\}$ necessary and sufficient

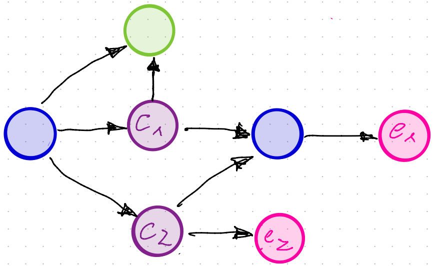
$\{c_1\}$ sufficient, not necessary

$\{c_2\}$ sufficient, not necessary

Example: necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E	iff	$\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$
C is a sufficient cause for E	iff	$\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$



effect set $E = \{e_1, e_2\}$

$\{c_1, c_2\}$ necessary, not sufficient

$\{c_1\}$ neither necessary nor sufficient

$\{c_2\}$ sufficient, not necessary

Pruning of necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

$$C \text{ is a necessary cause for } E \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$$

$$C \text{ is a sufficient cause for } E \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$$

If C is a necessary resp. sufficient cause for E then so is its pruning $\lfloor C \rfloor$, defined by:

$$\lfloor C \rfloor = \{ s \in C : \mathcal{M} \models \exists(\neg C) \, \mathrm{U} \, s \}$$

$\lfloor C \rfloor$ results from C by removing all states s where each path π from an initial state to s traverses another C-state. Hence: $\pi \models \Diamond C$ iff $\pi \models \Diamond \lfloor C \rfloor$.

Pruning of necessary and sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

$$C \text{ is a necessary cause for } E \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$$

$$C \text{ is a sufficient cause for } E \quad \text{iff} \quad \mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$$

If C is a necessary resp. sufficient cause for E then so is its pruning $\lfloor C \rfloor$, defined by:

$$\lfloor C \rfloor = \{ s \in C : \mathcal{M} \models \exists(\neg C) \,\mathsf{U}\, s \}$$

... towards small and early causes ("root causes") ...

Let's have a closer look: sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$

Let's have a closer look: sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$

Properties of sufficient causes:

- $C_E \overset{\text{def}}{=} \{s \in S : s \models \forall \bigcirc \forall \Diamond E\}$ is a sufficient cause

  ... and contains all other sufficient causes

Let's have a closer look: sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \to \Diamond E)$

Properties of sufficient causes:

- $C_E \overset{\text{def}}{=} \{s \in S : s \models \forall\bigcirc\forall\Diamond E\}$ is a sufficient cause
  ... and contains all other sufficient causes
- If $\underbrace{\mathcal{M} \not\models \exists\Diamond\forall\bigcirc\forall\Diamond E}_{\text{there is no reachable state } s \text{ s.t. } s \models \forall\bigcirc\forall\Diamond E}$ then $\varnothing$ is the only sufficient cause.

Let's have a closer look: sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \rightarrow \Diamond E)$

Properties of sufficient causes:

- $C_E \stackrel{\text{def}}{=} \{s \in S : s \models \forall\bigcirc\forall\Diamond E\}$ is a sufficient cause
  ... and contains all other sufficient causes
- If $\mathcal{M} \not\models \exists\Diamond\forall\bigcirc\forall\Diamond E$ then $\varnothing$ is the only sufficient cause.
- Canonical sufficient cause: $\lfloor C_E \rfloor$

pruning operator: $\lfloor C \rfloor = \{s \in C : \mathcal{M} \models \exists(\neg C) \,\mathsf{U}\, s\}$

Let's have a closer look: sufficient causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a sufficient cause for E iff $\mathcal{M} \models \forall(\Diamond C \to \Diamond E)$

Properties of sufficient causes:

- $C_E \stackrel{\text{def}}{=} \{s \in S : s \models \forall\bigcirc\forall\Diamond E\}$ is a sufficient cause
  ... and contains all other sufficient causes
- If $\mathcal{M} \not\models \exists\Diamond\forall\bigcirc\forall\Diamond E$ then $\varnothing$ is the only sufficient cause.
- Canonical sufficient cause: $\lfloor C_E \rfloor$

  ... is indeed a good one, with maximal degree of necessity (see later)

Let's have a closer look: necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$

Properties of necessary causes:

Let's have a closer look: necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$

Properties of necessary causes:

- The set I of initial states is a trivial necessary cause.

Let's have a closer look: necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \to \Diamond C)$

Properties of necessary causes:

- The set I of initial states is a trivial necessary cause.
- $Pre(E) = \{s : \exists s' \in E \text{ s.t. } s \to s'\}$ is a necessary cause for E.

Let's have a closer look: necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$. Then:

C is a necessary cause for E iff $\mathcal{M} \models \forall(\Diamond E \rightarrow \Diamond C)$

Properties of necessary causes:

- The set I of initial states is a trivial necessary cause.
- $Pre(E) = \{ s : \exists s' \in E \text{ s.t. } s \rightarrow s' \}$ is a necessary cause for E.
- How to define "good necessary causes"?

  Idea: seek for necessary causes that are "maximal sufficient"

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

degree of suffiency ("precision")

$$\Pr_{\mathcal{M}}(\Diamond E \mid \Diamond C) = \frac{\Pr_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond C)}$$

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

degree of suffiency ("precision")

$$\mathrm{Pr}_{\mathcal{M}}(\,\Diamond E \mid \Diamond C\,) \;=\; \frac{\mathrm{Pr}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\mathrm{Pr}_{\mathcal{M}}(\Diamond C)}$$

If C is a sufficient cause then the degree of sufficiency is 1.

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

degree of suffiency ("precision")

$$\Pr_{\mathcal{M}}(\,\Diamond E \mid \Diamond C\,) \;=\; \frac{\Pr_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond C)}$$

degree of necessity ("recall")

$$\Pr_{\mathcal{M}}(\,\Diamond C \mid \Diamond E\,) \;=\; \frac{\Pr_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond E)}$$

If C is a sufficient cause then the degree of sufficiency is 1.

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

degree of suffiency ("precision")
$$\Pr_{\mathcal{M}}(\ \Diamond E \mid \Diamond C\ ) \ = \ \frac{\Pr_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond C)}$$

degree of necessity ("recall")
$$\Pr_{\mathcal{M}}(\ \Diamond C \mid \Diamond E\ ) \ = \ \frac{\Pr_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond E)}$$

If C is a necessary cause then the degree of necessity is 1.

Degree of sufficiency and necessity

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal). Let $C \subseteq S$ s.t. $C \cap E = \varnothing$ and $C, E \neq \varnothing$.

Consider $\mathcal{M}$ as a Markov chain (uniform distributions for the initial states and the successors of every state).

degree of suffiency ("precision")

$$\mathrm{Pr}_{\mathcal{M}}(\,\Diamond E \mid \Diamond C\,) = \frac{\mathrm{Pr}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\mathrm{Pr}_{\mathcal{M}}(\Diamond C)}$$

degree of necessity ("recall")

$$\mathrm{Pr}_{\mathcal{M}}(\,\Diamond C \mid \Diamond E\,) = \frac{\mathrm{Pr}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\mathrm{Pr}_{\mathcal{M}}(\Diamond E)}$$

C and $\lfloor C \rfloor$ have the same degree of suffiency and necessity.

Optimal sufficient and necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal, nonempty).

Optimal sufficient and necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal, nonempty).

Sufficient cause with maximal degree of necessity:

$$\lfloor C_E \rfloor \quad \text{where} \quad C_E \stackrel{\text{def}}{=} \{ s \in S : s \models \forall \bigcirc \forall \Diamond E \}$$

Optimal sufficient and necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal, nonempty).

Sufficient cause with maximal degree of necessity:

$$\lfloor C_E \rfloor \quad \text{where} \quad C_E \stackrel{\text{def}}{=} \{ s \in S : s \models \forall \bigcirc \forall \Diamond E \}$$

Necessary causes with maximal degree of sufficiency:

$$\lfloor Pre(E) \rfloor$$

Optimal sufficient and necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal, nonempty).

Sufficient cause with maximal degree of necessity:

$$\lfloor C_E \rfloor \quad \text{where} \quad C_E \stackrel{\text{def}}{=} \{ s \in S : s \models \forall \bigcirc \forall \Diamond E \}$$

Necessary causes with maximal degree of sufficiency:

$$\lfloor Pre(E) \rfloor \quad \text{and} \quad \lfloor C \rfloor \text{ where } C = \{ s \in S : \Pr_s(\Diamond Pre(E)) = 1 \}$$

Optimal sufficient and necessary causes

Given a TS $\mathcal{M}$ with state space S and a set $E \subseteq S$ of effect states (non-initial, terminal, nonempty).

Sufficient cause with maximal degree of necessity:

$$\lfloor C_E \rfloor \quad \text{where} \quad C_E \stackrel{\text{def}}{=} \{ s \in S : s \models \forall \bigcirc \forall \Diamond E \}$$

Necessary causes with maximal degree of sufficiency:

$$\lfloor Pre(E) \rfloor \quad \text{and} \quad \lfloor C \rfloor \text{ where } C = \{ s \in S : \Pr_s(\Diamond Pre(E)) = 1 \}$$

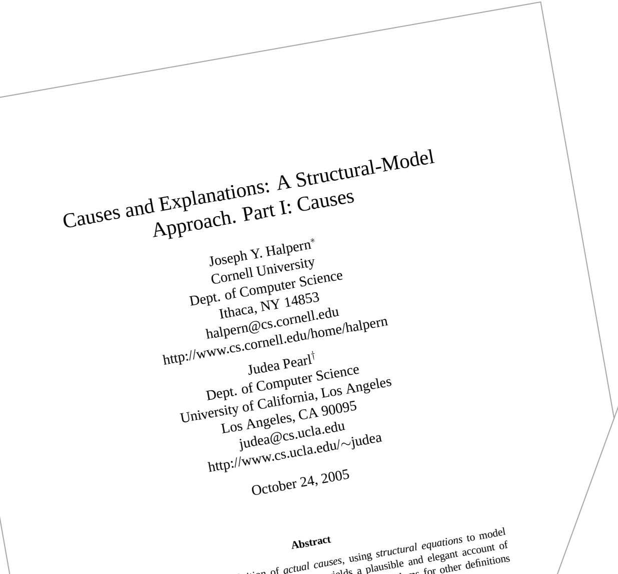
State-minimal necessary causes computable in polynomial time

using algorithms for weight-minimal s-t-cuts in directed graphs

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
  - Halpern-Pearl's approach to counterfactual causality
  - mutation-based forward responsibility
  - game-based forward and backward responsibility
  - quantitative responsibility via Shapley values
- Probabilistic causality in Markovian models
- Conclusions

Halpern-Pearl's approach to causality


Causes and Explanations: A Structural-Model Approach. Part I: Causes

Joseph Y. Halpern*
Cornell University
Dept. of Computer Science
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

Judea Pearl†
Dept. of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
judea@cs.ucla.edu
http://www.cs.ucla.edu/∼judea

October 24, 2005

Abstract


A Modification of the Halpern-Pearl Definition of Causality

Joseph Y. Halpern*
Cornell University
Computer Science Department
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

Abstract

The original Halpern-Pearl definition of causality [Halpern and Pearl, 2001] was updated in the journal version of the paper [Halpern and Pearl, 2005] to deal with some problems pointed out by Hopkins and Pearl [2003]. Here the definition is modified yet again, in a way that (a) leads to a simpler definition, (b) handles the problems pointed out by Hopkins and Pearl, and many others, (c) gives reasonable answers (that agree with those of the original and updated definition) in the standard problematic examples of causality, and (d) has lower complexity than either the original or updated definitions.

1 Introduction

Causality plays a central role in the way people structure the world. People constantly seek causal explanations for their observations. Philosophers have typically distinguished two notions of causality, which they have called *type causality* (sometimes called *general causality*) and *actual causality* (sometimes called *token causality* or *specific causality*). Type causality is perhaps what scientists are most concerned with. These are general statements, such as "smoking causes lung cancer" and "printing money causes inflation". By

However, as is well known, the but-for test is not always sufficient to determine causality. Consider the following well-known example, taken from [Paul and Hall, 2013]:

Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

Here the but-for test fails. Even if Suzy hadn't thrown, the bottle would have shattered. Nevertheless, we want to call Suzy's throw a cause of the bottle shattering.

Halpern and Pearl [2001] introduced a definition using *structural equations* that has proved quite influential. In the structural-equations approach, the world is assumed to be characterized by the values of a collection of variables. In this example, we can use binary variable ST for "Suzy throws" ($ST = 1$ if Suzy throws; $ST = 0$ if she doesn't), BT for "Billy throws", and BS for "bottle shatters".

Halpern-Pearl's approach to causality

- actual/specific vs general/type causes

  actual cause is a factual event C that causes the effect E

  general cause: e.g. "sweets cause obesity"

- backward vs forward causality-based reasoning

  backward: what has caused an observed effect E (e.g., observed event sequence)?

  forward: what can cause an event E in a given world model?

- counterfactual vs necessary vs sufficient cause-effect relations

  counterfactual: if C would not have happened, then E would not have occured

  necessary: if E occurs then C must have happened before

  sufficient: if C happens then always E will occur somewhen later

- deterministic vs probabilistic causes, and many more ...

HP structural equation model

HP structural equation model

Structural equation model: $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ where

Exo: set of exogenous variables (specify the context)

Endo: totally ordered set of endogenous variables, say $x_1, \ldots, x_n$

x_1 only depends on the context
x_2 only depends on the context and x_1
x_3 only depends on the context and x_1, x_2
$\vdots \qquad \vdots$

HP structural equation model

Structural equation model: $\mathcal{S} = (Exo, Endo, f)$ where

- Exo: set of exogenous variables (specify the context)
- $Endo$: totally ordered set of endogenous variables, say $x_1, \ldots, x_n$
- $f = (f_1, \ldots, f_n)$ where $f_i : Val(Exo, x_1, \ldots, x_{i-1}) \to Val(x_i)$

f yields the values of the endo variables for context $c \in Val(Exo)$

x_1 only depends on the context
x_2 only depends on the context and x_1
$\vdots$ $\vdots$

$Val(\mathcal{V})$ = set of valuations for the variables in $\mathcal{V} \subseteq Exo \cup Endo$

HP structural equation model

Structural equation model: $\mathcal{S} = (Exo, Endo, f)$ where

Exo: set of exogenous variables (specify the context)

$Endo$: totally ordered set of endogenous variables, say $x_1, \ldots, x_n$

$f = (f_1, \ldots, f_n)$ where $f_i : Val(Exo, x_1, \ldots, x_{i-1}) \to Val(x_i)$

f yields the values of the endo variables for context $c \in Val(Exo)$:

$$
\begin{array}{lllll}
\alpha_1 & = & \mathcal{S}_1(c) & \stackrel{\text{def}}{=} & f_1(c) & \text{(value for } x_1\text{)} \\
\alpha_2 & = & \mathcal{S}_2(c) & \stackrel{\text{def}}{=} & f_2(c, \alpha_1) & \text{(value for } x_2\text{)} \\
 & & \vdots & & \vdots & \vdots \\
\alpha_n & = & \mathcal{S}_n(c) & \stackrel{\text{def}}{=} & f_n(c, \alpha_1, \ldots, \alpha_{n-1}) & \text{(value for } x_n\text{)}
\end{array}
$$

HP structural equation model

Structural equation model: $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ where

Exo: set of exogenous variables (specify the context)

Endo: totally ordered set of endogenous variables, say $x_1, \ldots, x_n$

$f = (f_1, \ldots, f_n)$ where $f_i : \mathit{Val}(\mathit{Exo}, x_1, \ldots, x_{i-1}) \to \mathit{Val}(x_i)$

Interventions:

for counterfactual reasoning:

"enforce values of endogenous variables (ignoring their equations)"

HP structural equation model

Structural equation model: $\mathcal{S} = (Exo, Endo, f)$ where

Exo: set of exogenous variables (specify the context)

$Endo$: totally ordered set of endogenous variables, say $x_1, \ldots, x_n$

$f = (f_1, \ldots, f_n)$ where $f_i : Val(Exo, x_1, \ldots, x_{i-1}) \to Val(x_i)$

Interventions: given $Y \subseteq Endo$ and $\beta \in Val(Y)$, let

$$\mathcal{S}[Y \leftarrow \beta] = \left\{ \begin{array}{l} \mathcal{S} \text{ when the } Y\text{-variables are treated as} \\ \text{constants given by the values in } \beta \end{array} \right.$$

for counterfactual reasoning:

"enforce values of endogenous variables (ignoring their equations)"

HP causality

HP causality

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$

HP causality

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $X \subseteq \mathit{Endo}$ and $\alpha = \mathcal{S}_X(c)$

tuple of values for X
in $\mathcal{S}$ for context c
obtained by the equations
$x_i = f_i(c, x_1, \ldots, x_{i-1})$

HP causality

Let $\mathcal{S} = (Exo, Endo, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in Val(Exo)$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $X \subseteq Endo$ and $\alpha = \mathcal{S}_X(c)$

Then $X{=}\alpha$ is called a cause for φ in context c iff

[AC1] ... counterfactual condition ...

[AC2] ... minimality condition ....

HP causality

Let $\mathcal{S} = (Exo, Endo, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in Val(Exo)$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $X \subseteq Endo$ and $\alpha = \mathcal{S}_X(c)$

Then $X{=}\alpha$ is a but-for cause for φ in context c iff

[AC1] There is $\beta \in Val(X)$ such that

$$(\mathcal{S}[X{\leftarrow}\beta], c) \models \neg\varphi$$

[AC2] ... minimality condition ....

HP causality

[Halpern'15]

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $X \subseteq \mathit{Endo}$ and $\alpha = \mathcal{S}_X(c)$

Then $X{=}\alpha$ is an actual cause for φ in context c iff

[AC1] There is $\beta \in \mathit{Val}(X)$ and $Y \subseteq \mathit{Endo}$ such that

$$(\mathcal{S}[X{\leftarrow}\beta, Y{\leftarrow}\mathcal{S}_Y(c)], c) \models \neg\varphi$$

[AC2] ... minimality condition ....

HP causality

[Halpern'15]

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $X \subseteq \mathit{Endo}$ and $\alpha = \mathcal{S}_X(c)$

Then $X{=}\alpha$ is an actual cause for φ in context c iff

[AC1] There is $\beta \in \mathit{Val}(X)$ and $Y \subseteq \mathit{Endo}$ such that

$$(\mathcal{S}[X{\leftarrow}\beta, Y{\leftarrow}\mathcal{S}_Y(c)], c) \models \neg\varphi$$

[AC2] X is minimal w.r.t. condition [AC1]

HP causality and degree of responsibility

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $x \in \mathit{Endo}$ and $\alpha = \mathcal{S}_x(c)$

Then, the degree of responsibility of $x{=}\alpha$ for φ is . . .

[Chockler/Halpern/Kupferman, ACM ToCL 2008]

HP causality and degree of responsibility

Let $\mathcal{S} = (Exo, Endo, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in Val(Exo)$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $x \in Endo$ and $\alpha = \mathcal{S}_x(c)$

Then, the degree of responsibility of $x{=}\alpha$ for φ is $\frac{1}{m}$ where

$$m = \begin{cases} \text{minimal number of value-changes for endo variables} \\ \text{required to make } \varphi \text{ counterfactually depend on } x \end{cases}$$

[Chockler/Halpern/Kupferman, ACM ToCL 2008]

HP causality and degree of responsibility

Let $\mathcal{S} = (\mathit{Exo}, \mathit{Endo}, f)$ be a structural equation model and

- φ be a Boolean conditon for the values of variables (exo or endo)
- $c \in \mathit{Val}(\mathit{Exo})$ a context s.t. $(\mathcal{S}, c) \models \varphi$
- $x \in \mathit{Endo}$ and $\alpha = \mathcal{S}_x(c)$

Then, the degree of responsibility of $x{=}\alpha$ for φ is $\dfrac{1}{m}$ where

$$m = \begin{cases} \text{minimal number of value-changes for endo variables} \\ \text{required to make } \varphi \text{ counterfactually depend on } x \end{cases}$$

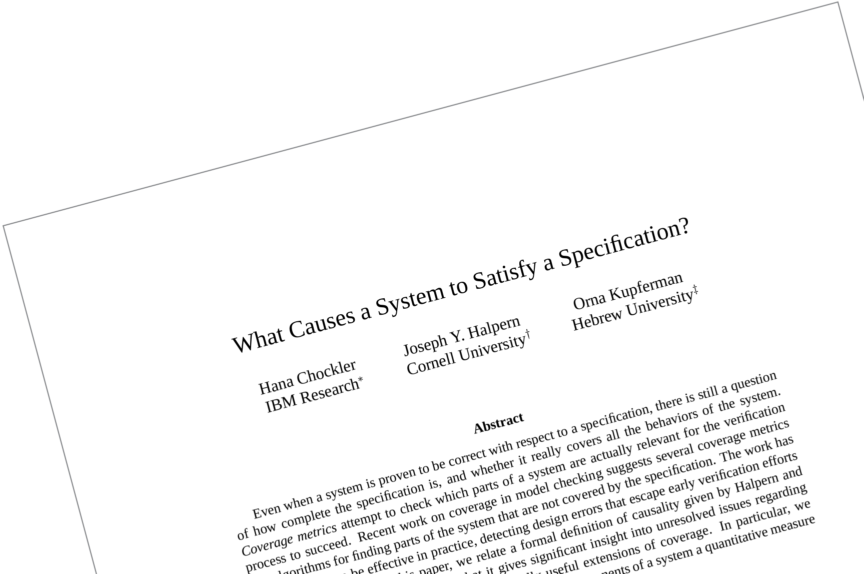
Formally: $m = |X|$ where X is a smallest set of endogenous variables that contains x and satisfies [AC1], i.e., there exist a valuation β for X and $Y \subseteq \mathit{Endo}$ s.t.:

$$(\mathcal{S}[X{\leftarrow}\beta, Y{\leftarrow}\mathcal{S}_Y(c)], c) \models \neg\varphi$$

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
    - Halpern-Pearl's approach to counterfactual causality
    - mutation-based forward responsibility
    - game-based forward and backward responsibility
    - quantitative responsibility via Shapley values
- Probabilistic causality in Markovian models
- Conclusions

HP-based responsibility in TS



What Causes a System to Satisfy a Specification?

Hana Chockler
IBM Research*

Joseph Y. Halpern
Cornell University†

Orna Kupferman
Hebrew University‡

Abstract

Even when a system is proven to be correct with respect to a specification, there is still a question of how complete the specification is, and whether it really covers all the behaviors of the system. *Coverage metrics* attempt to check which parts of a system are actually relevant for the verification process to succeed. Recent work on coverage in model checking suggests several coverage metrics and algorithms for finding parts of the system that are not covered by the specification. The work has ... be effective in practice, detecting design errors that escape early verification efforts ... this paper, we relate a formal definition of causality given by Halpern and ... it gives significant insight into unresolved issues regarding ... useful extensions of coverage. In particular, we ... of a system a quantitative measure

Counterfactuality: backward vs forward

Counterfactuality: backward vs forward

backward counterfactual causality

given an effect secanrio:

“if $\underbrace{\textit{the cause would not have happened}}_{\substack{\text{intervention:}\\ \text{modify cause items}}}$*, then the effect would not have occured”*

Counterfactuality: backward vs forward

backward counterfactual causality

given an effect secanrio:

“if $\underbrace{\textit{the cause would not have happened}}_{\text{intervention: modify cause items}}$*, then the effect would not have occured”*

forward counterfactual causality

given a world model:

“minimal set of items that need to be modified to avoid the effect”

Counterfactuality: backward vs forward

backward counterfactual causality

given an effect secanrio:

“$\underbrace{\textit{if the cause would not have happened,}}_{\text{intervention: modify cause items}}$ *then the effect would not have occured*”

forward counterfactual causality $=$ forward responsibility

given a world model:

“minimal set of items that need to be modified to avoid the effect”

degree of responsibility:

numerical values for individual cause items

HP-like causality and responsibility in TS

[Chockler/Halpern/Kupferman, ACM ToCL 2008]

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \to \{0, 1\}$.

Intuitively: $L_s(q) = 1$ iff atomic proposition q holds in state s

[Chockler/Halpern/Kupferman, ACM ToCL 2008]

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \to \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.

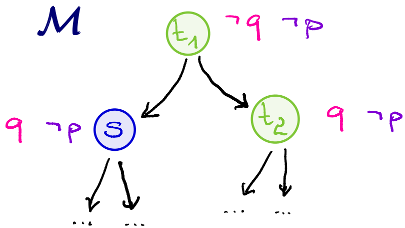
[Chockler/Halpern/Kupferman, ACM ToCL 2008]

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \to \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.



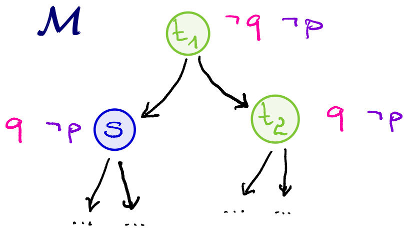
$T = \{t_1, t_2\}$

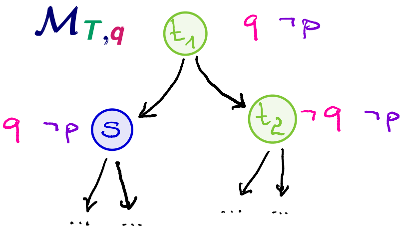
HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \rightarrow \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.





HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s\in S}$ where $L_s : AP \to \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.

Suppose $\mathcal{M} \models \phi$ (temporal property over 2^{AP}) and let $q \in AP$.

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s\in S}$ where $L_s : AP \to \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.

Suppose $\mathcal{M} \models \phi$ (temporal property over 2^{AP}) and let $q \in AP$.

- switching pair: (T, s) where $T \subseteq S$, $s \in S$ s.t.

$$\mathcal{M}_{T,q} \models \phi \quad \text{and} \quad \mathcal{M}_{T\cup\{s\},q} \not\models \phi$$

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s\in S}$ where $L_s : AP \to \{0,1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.

Suppose $\mathcal{M} \models \phi$ (temporal property over 2^{AP}) and let $q \in AP$.

- switching pair: (T, s) where $T \subseteq S$, $s \in S$ s.t.

$$\mathcal{M}_{T,q} \models \phi \quad \text{and} \quad \mathcal{M}_{T\cup\{s\},q} \not\models \phi$$

- state s is a q-cause state for $\mathcal{M} \models \phi$ if there exists a switching pair (T, s)

HP-like causality and responsibility in TS

Given a transition system $\mathcal{M}$ with state space S and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \to \{0, 1\}$.

Intervention ("mutations of the truth values of atomic propositions"):

- Given $q \in AP$ and $T \subseteq S$, then $\mathcal{M}_{T,q}$ is $\mathcal{M}$ with flipped labeling values $L_t(q)$ for $t \in T$.

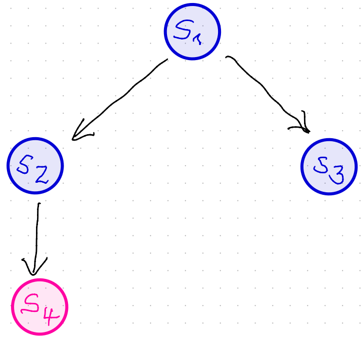
Suppose $\mathcal{M} \models \phi$ (temporal property over 2^{AP}) and let $q \in AP$.

- switching pair: (T, s) where $T \subseteq S$, $s \in S$ s.t.

$$\mathcal{M}_{T,q} \models \phi \quad \text{and} \quad \mathcal{M}_{T \cup \{s\},q} \not\models \phi$$

- degree of q-responsibility of cause state s is $1/(|T|+1)$ where (T, s) is a switching pair of minimal size

Example: responsibility à la Chockler et al



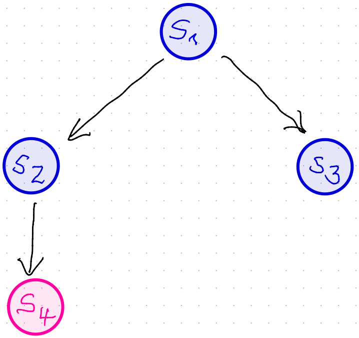
$AP = \{q\}$

$s_1, s_2, s_3 \not\models q$

$s_4 \models q$

$\mathcal{M} \models \exists\Diamond q$

Example: responsibility à la Chockler et al

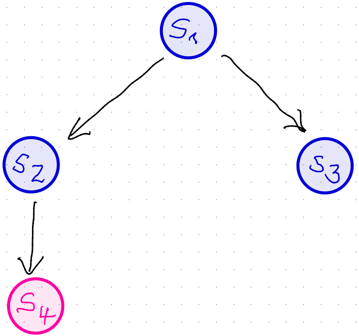


$AP = \{q\}$
$s_1, s_2, s_3 \not\models q$
$s_4 \models q$

$\mathcal{M} \models \exists\Diamond q$

$\mathcal{M}_{T,q} \not\models \exists\Diamond q$ iff $T = \{s_4\}$

Example: responsibility à la Chockler et al



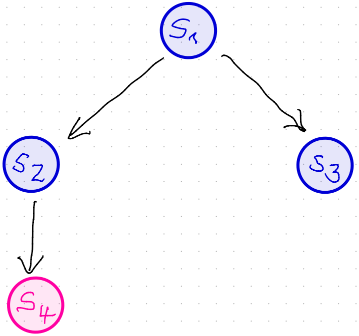
$AP = \{q\}$
$s_1, s_2, s_3 \not\models q$
$s_4 \models q$

$\mathcal{M} \models \exists\Diamond q$

$\mathcal{M}_{T,q} \not\models \exists\Diamond q$ iff $T = \{s_4\}$

$(\varnothing, s_4)$ is the only switching pair

Example: responsibility à la Chockler et al



$AP = \{q\}$
$s_1, s_2, s_3 \not\models q$
$s_4 \models q$

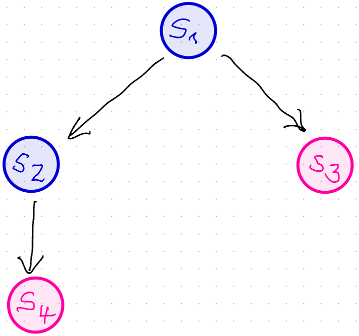
$\mathcal{M} \models \exists\Diamond q$

$\mathcal{M}_{T,q} \not\models \exists\Diamond q$ iff $T = \{s_4\}$

$(\varnothing, s_4)$ is the only switching pair

- s_4 is a q-cause state and has responsibility 1
- s_1, s_2, s_3 are not q-cause states and have responsibility 0

Example: responsibility à la Chockler et al



$AP = \{q\}$
$s_1, s_2 \not\models q$
$s_3, s_4 \models q$

$\mathcal{M} \models \exists\Diamond q$

$\mathcal{M}_{T,q} \not\models \exists\Diamond q$ iff $T = \{s_3, s_4\}$

2 switching pairs

- s_3, s_4 are q-cause states and have responsibility $1/2$
- s_1, s_2 are not q-cause states and have responsibility 0

HP-like causality and responsibility in TS

So far: notions of q-cause and degree of q-responsibility for fixed atomic proposition q

HP-like causality and responsibility in TS

So far: notions of q-cause and degree of q-responsibility for fixed atomic proposition q

Analogous definition independent of specific atomic proposition

Intervention:

- given $T \subseteq S \times AP$, then $\mathcal{M}_T$ equals $\mathcal{M}$ with flipped values for the pairs $(s, q) \in T$

HP-like causality and responsibility in TS

So far: notions of q-cause and degree of q-responsibility for fixed atomic proposition q

Analogous definition independent of specific atomic proposition

Intervention:

- given $T \subseteq S \times AP$, then $\mathcal{M}_T$ equals $\mathcal{M}$ with flipped values for the pairs $(s, q) \in T$

Suppose $\mathcal{M} \models \phi$

cause: set T s.t. $\mathcal{M}_T \not\models \phi$ and $\mathcal{M}_U \models \phi$ for any subset U of T

degree of responsibility of pair (s, q) is $1/(|T|+1)$ where $T \cup \{(s, q)\}$ is a cause of minimal size (under all causes containing (s, q))

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
  - Halpern-Pearl's approach to counterfactual causality
  - mutation-based forward responsibility
  - game-based forward and backward responsibility
  - quantitative responsibility via Shapley values
- Probabilistic causality in Markovian models
- Conclusions

Responsibility w.r.t. nondeterministic choices

[Baier/Funke/Majumdar, IJCAI'21]

Responsibility w.r.t. nondeterministic choices

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

[Baier/Funke/Majumdar, IJCAI'21]

Responsibility w.r.t. nondeterministic choices

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

- *forward:* in which states do we need to control the nondeterminism to ensure that ϕ does not hold in $\mathcal{M}$?

[Baier/Funke/Majumdar, IJCAI'21]

Responsibility w.r.t. nondeterministic choices

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

- *forward:* in which states do we need to control the nondeterminism to ensure that ϕ does not hold in $\mathcal{M}$?
- *backward:* for a given execution where ϕ holds, which states were responsible for the satisfaction of ϕ?

  which states would have had the option to avoid the bad event by resolving the nondeterministic choices in a different way?

[Baier/Funke/Majumdar, IJCAI'21]

Responsibility w.r.t. nondeterministic choices

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Game-based notions of responsibility for sets $C \subseteq S$

w.r.t. to their power of avoiding the bad event in terms of their nondeterministic choices

[Baier/Funke/Majumdar, IJCAI'21]

Responsibility w.r.t. nondeterministic choices

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Game-based notions of responsibility for sets $C \subseteq S$

> w.r.t. to their power of avoiding the bad event in terms of their nondeterministic choices

using the two-player game structure $\mathcal{M}_C$:

- arena: state space, initial state and transitions of $\mathcal{M}$
- player 1 controls all states in C (objective $\neg\phi$)
- player 2 controls all states in $\overline{C} = S \setminus C$ (objective ϕ)

Forward responsibility for temporal properties

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Let $C \subseteq S$. Then, C is forward responsible for ϕ if

Forward responsibility for temporal properties

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Let $C \subseteq S$. Then, C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

i.e., a strategy σ for player 1 s.t. the bad event does not happen in σ-plays

Forward responsibility for temporal properties

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Let $C \subseteq S$. Then, C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

i.e., a strategy σ for player 1 s.t. the bad event does not happen in σ-plays

[F2] C is minimal w.r.t. [F1]

i.e., no proper subset can ensure that the bad event does not happen

Forward responsibility for temporal properties

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Let $C \subseteq S$. Then, C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

i.e., a strategy σ for player 1 s.t. the bad event does not happen in σ-plays

[F2] C is minimal w.r.t. [F1]

i.e., no proper subset can ensure that the bad event does not happen

Observations:

- If $\mathcal{M} \models \forall\phi$ then noone is forward responsible, and vice versa.

Forward responsibility for temporal properties

Starting point: transition system $\mathcal{M}$ with state space S and a path property ϕ (bad event).

Let $C \subseteq S$. Then, C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

i.e., a strategy σ for player 1 s.t. the bad event does not happen in σ-plays

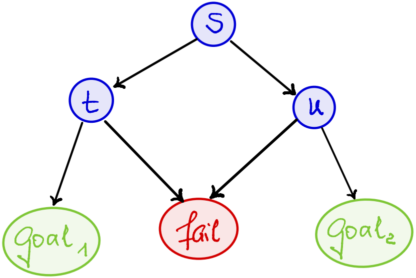
[F2] C is minimal w.r.t. [F1]

i.e., no proper subset can ensure that the bad event does not happen

Observations:

- If $\mathcal{M} \models \forall\phi$ then noone is forward responsible, and vice versa.
- If $\mathcal{M} \models \forall\neg\phi$ then exactly $C = \varnothing$ is forward responsible.

Forward responsibility: example



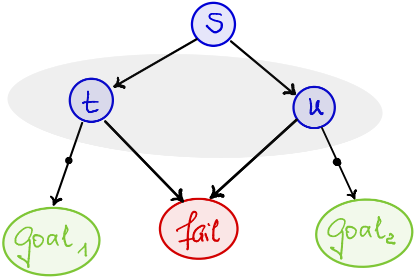
$\phi = \Diamond \textit{fail}$ ("bad event")

C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

[F2] C is minimal w.r.t. [F1]

Forward responsibility: example



$\phi = \Diamond \textit{fail}$ ("bad event")

forward responsible sets:

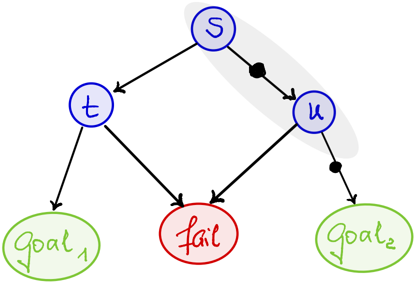
$\{t, u\}$

C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

[F2] C is minimal w.r.t. [F1]

Forward responsibility: example



$\phi = \Diamond \textit{fail}$ ("bad event")

forward responsible sets:

$\{t, u\}$

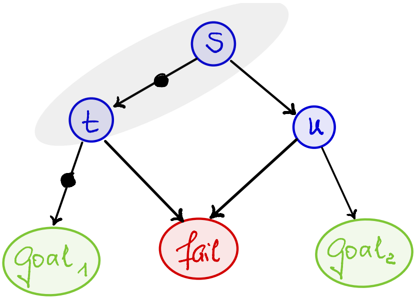
$\{s, u\}$

C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

[F2] C is minimal w.r.t. [F1]

Forward responsibility: example



$\phi = \lozenge\, \mathit{fail}$ ("bad event")

forward responsible sets:

$\{t, u\}$

$\{s, u\}$

$\{s, t\}$

C is forward responsible for ϕ if

[F1] C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$

[F2] C is minimal w.r.t. [F1]

Responsibility in TS

- so far: forward responsibility

  "which states are responsible for the satisfaction of a property of the entire model?"

- now: backward responsibility

  "which states are responsible for the satisfaction of an undesired property along a given error scenario?"

Responsibility in TS

- so far: forward responsibility

  "which states are responsible for the satisfaction of a property of the entire model?"

- now: backward responsibility

  "which states are responsible for the satisfaction of an undesired property along a given error scenario?"

  - ⋆ strategic view: error scenario is a path
  - ⋆ causality-based view: error scenario is a path + strategy for opponents

Strategic backward responsibility

Strategic backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a path $\pi = s_0\, s_1\, s_2 \ldots$ s.t. $\pi \models \phi$.

Strategic backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a path $\pi = s_0\, s_1\, s_2 \ldots$ s.t. $\pi \models \phi$.

C is strategically backward responsible for "$\pi \models \phi$" if

[SB1] there exists $n \in \mathbb{N}$ such that C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$ from state s_n

i.e., C could have played differently from s_n to enforce the violation of ϕ

Strategic backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a path $\pi = s_0\, s_1\, s_2 \ldots$ s.t. $\pi \models \phi$.

C is strategically backward responsible for "$\pi \models \phi$" if

[SB1] there exists $n \in \mathbb{N}$ such that C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$ from state s_n

i.e., C could have played differently from s_n to enforce the violation of ϕ

[SB2] C is minimal w.r.t. [SB1]

Strategic backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a path $\pi = s_0\, s_1\, s_2 \ldots$ s.t. $\pi \models \phi$.

C is strategically backward responsible for "$\pi \models \phi$" if

[SB1] there exists $n \in \mathbb{N}$ such that C has a winning strategy in $\mathcal{M}_C$ for objective $\neg\phi$ from state s_n

i.e., C could have played differently from s_n to enforce the violation of ϕ

[SB2] C is minimal w.r.t. [SB1]

---

objective from state s_n: $\neg\phi$ if ϕ is prefix independent,

but residual property "$\neg\phi$ after $s_0 \ldots s_{n-1}$" in the general case

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$

Strategy profile σ specifies

- a path (the unique σ-play π_σ)
- $\overline{C}$'s decision along other paths (for counterfactual reasoning)
- C's decision along other paths (irrelevant)

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$ s.t. $\underbrace{\mathcal{M}, \sigma \models \phi}_{\substack{\pi_\sigma \models \phi \\ \text{for the unique} \\ \sigma\text{-play } \pi_\sigma}}$.

Strategy profile σ specifies

- a path (the unique σ-play π_σ)
- $\overline{C}$'s decision along other paths (for counterfactual reasoning)
- C's decision along other paths (irrelevant)

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$ s.t. $\mathcal{M}, \sigma \models \phi$.

C is causally backward responsible for "$\mathcal{M}, \sigma \models \phi$" if

[CB1] there exists a strategy τ_C for C in $\mathcal{M}_C$ s.t. the unique $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

Strategy profile σ specifies

- a path (the unique σ-play π_σ)
- $\overline{C}$'s decision along other paths (for counterfactual reasoning)
- C's decision along other paths (irrelevant)

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$ s.t. $\mathcal{M}, \sigma \models \phi$.

C is causally backward responsible for "$\mathcal{M}, \sigma \models \phi$" if

[CB1] there exists a strategy τ_C for C in $\mathcal{M}_C$ s.t. the unique $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

i.e., C could have played differently to enforce the violation of ϕ, when the strategy for the other states is fixed

Causal backward responsibility

Given TS $\mathcal{M}$, path property ϕ, a set C of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$ s.t. $\mathcal{M}, \sigma \models \phi$.

C is causally backward responsible for "$\mathcal{M}, \sigma \models \phi$" if

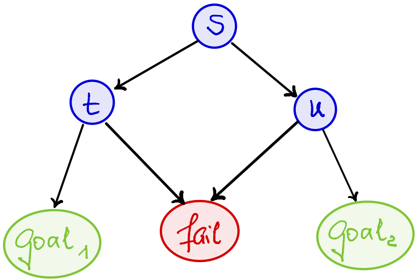
[CB1] there exists a strategy τ_C for C in $\mathcal{M}_C$ s.t. the unique $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

i.e., C could have played differently to enforce the violation of ϕ, when the strategy for the other states is fixed

[CB2] C is minimal w.r.t. [CB1]

i.e., no proper subset of C can enforce the violation of ϕ, when the other states stick to their strategy

Backward responsibility: example (strategic)



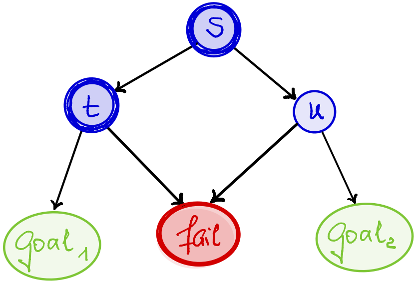
$\phi = \Diamond \textit{fail}$ ("bad event")

C is strategically backward responsible for $s_0\, s_1\, s_2 \ldots \models \phi$ if

[SB1] there is n s.t. C has a winning strategy for $\neg\phi$ from s_n

[SB2] C is minimal w.r.t. [SB1]

Backward responsibility: example (strategic)



$\phi = \Diamond \textit{fail}$ ("bad event")

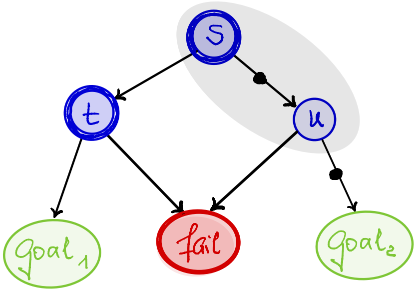
path $s\ t\ \textit{fail} \models \phi$

C is strategically backward responsible for $s_0\, s_1\, s_2 \ldots \models \phi$ if

[SB1] there is n s.t. C has a winning strategy for $\neg\phi$ from s_n

[SB2] C is minimal w.r.t. [SB1]

Backward responsibility: example (strategic)



$\phi = \Diamond \textit{fail}$ ("bad event")

path $s\, t\, \textit{fail} \models \phi$

strat-backward responsible:

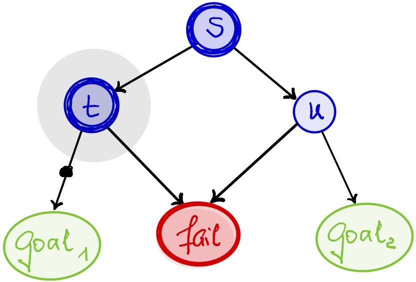
$\{s, u\}$

C is strategically backward responsible for $s_0\, s_1\, s_2 \ldots \models \phi$ if

[SB1] there is n s.t. C has a winning strategy for $\neg\phi$ from s_n

[SB2] C is minimal w.r.t. [SB1]

Backward responsibility: example (strategic)



$\phi = \Diamond \textit{fail}$ ("bad event")

path $s\ t\ \textit{fail} \models \phi$

strat-backward responsible:

$\{s, u\}$

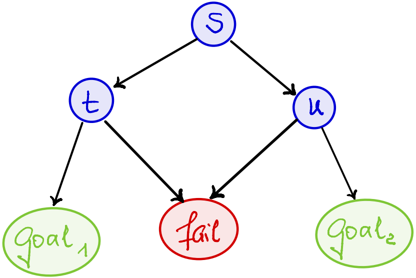
$\{t\}$

C is strategically backward responsible for $s_0\, s_1\, s_2 \ldots \models \phi$ if

[SB1] there is n s.t. C has a winning strategy for $\neg\phi$ from s_n

[SB2] C is minimal w.r.t. [SB1]

Backward responsibility: example (causal)



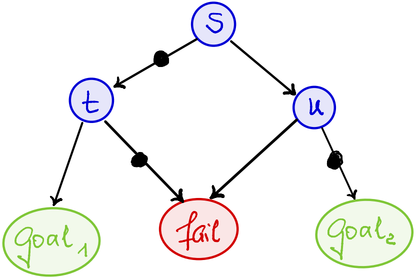
$\phi = \Diamond$***fail*** ("bad event")

C is causally backward responsible for $(\sigma_C, \sigma_{\overline{C}}) \models \phi$ if

[CB1] there is a strategy τ_C for C s.t. the $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

[CB2] C is minimal w.r.t. [CB1]

Backward responsibility: example (causal)



$\phi = \Diamond \textit{fail}$ ("bad event")

strategy profile:

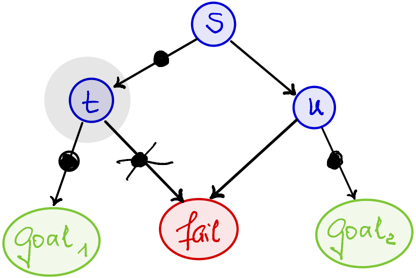
$s \to t, \ t \to f, \ u \to g_2$

C is causally backward responsible for $(\sigma_C, \sigma_{\overline{C}}) \models \phi$ if

[CB1] there is a strategy τ_C for C s.t. the $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

[CB2] C is minimal w.r.t. [CB1]

Backward responsibility: example (causal)



$\phi = \Diamond \mathit{fail}$ ("bad event")

strategy profile:

$s \to t$, $t \to f$, $u \to g_2$

causally backward responsible:

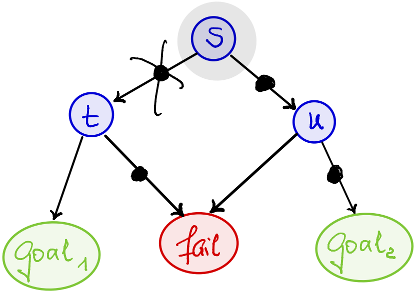
$\{t\}$; change $t \to g_1$

C is causally backward responsible for $(\sigma_C, \sigma_{\overline{C}}) \models \phi$ if

[CB1] there is a strategy τ_C for C s.t. the $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

[CB2] C is minimal w.r.t. [CB1]

Backward responsibility: example (causal)



$\phi = \Diamond \mathit{fail}$ ("bad event")

strategy profile:

$s \to t$, $t \to f$, $u \to g_2$

causally backward responsible:

$\{t\}$; change $t \to g_1$

$\{s\}$; change $s \to u$

C is causally backward responsible for $(\sigma_C, \sigma_{\overline{C}}) \models \phi$ if

[CB1] there is a strategy τ_C for C s.t. the $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg\phi$

[CB2] C is minimal w.r.t. [CB1]

Relation between f-, sb- and cb-responsibility

f-responsible	=	forward responsible
sb-responsible	=	strategically backward responsible
cb-responsible	=	causally backward responsible

Relation between f-, sb- and cb-responsibility

$$\text{f-responsibility} \Longrightarrow \text{sb-responsibility} \Longrightarrow \text{cb-responsibility}$$

up to minimality | up to minimality

f-responsible	=	forward responsible
sb-responsible	=	strategically backward responsible
cb-responsible	=	causally backward responsible

Relation between f-, sb- and cb-responsibility

f-responsibility $\implies$ sb-responsibility $\implies$ cb-responsibility

Let C be a set of states.

- C is f-responsible for ϕ iff C contains a coalition that is sb-responsible for all $\pi \models \phi$, and is minimal w.r.t. this property.

f-responsible = forward responsible
sb-responsible = strategically backward responsible
cb-responsible = causally backward responsible

Relation between f-, sb- and cb-responsibility

f-responsibility $\Longrightarrow$ sb-responsibility $\Longrightarrow$ cb-responsibility

Let C be a set of states.

- C is f-responsible for ϕ iff C contains a coalition that is sb-responsible for all $\pi \models \phi$, and is minimal w.r.t. this property.
- If C is sb-responsible for $\pi \models \phi$ and σ a strategy profile s.t. π is the σ-play then C contains a coalition that is cb-responsible for $\mathcal{M}, \sigma \models \phi$.

Relation between f-, sb- and cb-responsibility

f-responsibility $\Longrightarrow$ sb-responsibility $\Longrightarrow$ cb-responsibility

generalizes
HP-causality in SEM

Let C be a set of states.

- C is f-responsible for ϕ iff C cont
  sb-responsible for all $\pi \models \phi$, and
  property.

- If C is sb-responsible for $\pi \models \phi$ and σ a strategy profile s.t. π is the σ-play then C contains a coalition that is cb-responsible for $\mathcal{M}, \sigma \models \phi$.

HP-causality and cb-responsibility

structural equation model $\mathcal{S} = (Exo, Endo, f)$
context $c \in Val(Exo)$

$\Downarrow$

tree-like transition system $\mathcal{M}_{\mathcal{S},c}$

HP-causality and cb-responsibility

structural equation model $\mathcal{S} = (Exo, Endo, f)$
context $c \in Val(Exo)$

$\Downarrow$

tree-like transition system $\mathcal{M}_{\mathcal{S},c}$

total order for endo variables: $x_1, \ldots, x_n$

- root (level 0): given context c
- states at level $i \in \{1, \ldots, n\}$: valuations for $x_1, \ldots, x_{i-1}, x_i$
- transitions of state $s = [x_1{=}\alpha_1, \ldots, x_{i-1}{=}\alpha_{i-1}]$ at level $i-1$:

  default transition: $s \to [s, x_i{=}f_i(c, s)]$

  intervention: $s \to [s, x_i{=}\beta]$ for any other value β

HP-causality and cb-responsibility

structural equation model $\mathcal{S} = (Exo, Endo, f)$
context $c \in Val(Exo)$

$\Downarrow$

tree-like transition system $\mathcal{M}_{\mathcal{S},c}$

total order for endo variables: $x_1, \ldots, x_n$

Given a Boolean condition φ for the endogenous variables:

$X{=}\alpha$ is a but-for cause for φ

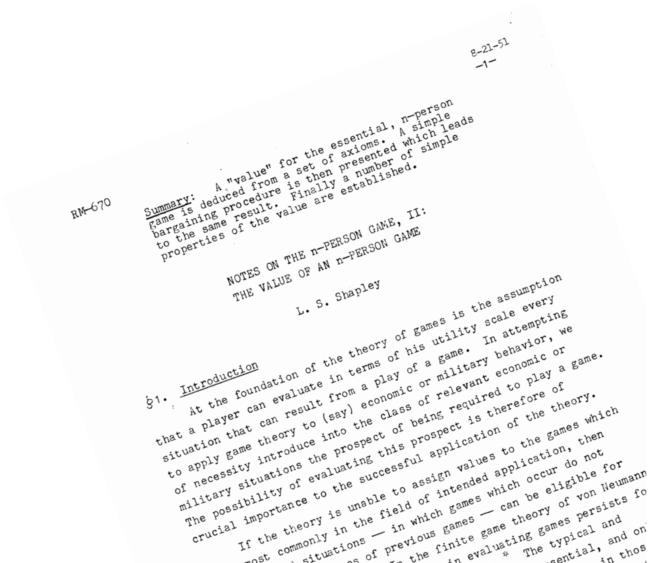
iff the X-states constitute a cb-responsible coalition for ϕ under the default strategy profile

where $\phi = \Diamond$ "φ *holds at some leave*" and $\alpha = \mathcal{S}_X(c)$

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
  - Halpern-Pearl's approach to counterfactual causality
  - mutation-based forward responsibility
  - game-based forward and backward responsibility
  - quantitative responsibility via Shapley values
- Probabilistic causality in Markovian models
- Conclusions

Shapley values


8-21-51
-1-

RM-670

Summary: A "value" for the essential, n-person game is deduced from a set of axioms. A simple bargaining procedure is then presented which leads to the same result. Finally a number of simple properties of the value are established.

NOTES ON THE n-PERSON GAME, II:
THE VALUE OF AN n-PERSON GAME

L. S. Shapley

§1. Introduction

At the foundation of the theory of games is the assumption that a player can evaluate in terms of his utility scale every situation that can result from a play of a game. In attempting to apply game theory to (say) economic or military behavior, we of necessity introduce into the class of relevant economic or military situations the prospect of being required to play a game. The possibility of evaluating this prospect is therefore of crucial importance to the successful application of the theory.

If the theory is unable to assign values to the games which

Lloyd S. Shapley

(Nobel prize 2012 for Economics)

Cooperative games and Shapley values

Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $Ag = \{1, \ldots, n\}$,
- a payoff function $val : 2^{Ag} \to \mathbb{R}$ s.t. $val(\varnothing) = 0$

$$val(C) = \text{value of coalition } C \subseteq Ag$$

Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $Ag = \{1, \ldots, n\}$,
- a payoff function $val : 2^{Ag} \to \mathbb{R}$ s.t. $val(\varnothing) = 0$

Given a total order π of Ag and an agent $a \in Ag$:

$$\pi_{\geqslant a} = \{i \in Ag \mid \pi(i) \geqslant \pi(a)\}$$

Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $Ag = \{1, \ldots, n\}$,
- a payoff function $val : 2^{Ag} \to \mathbb{R}$ s.t. $val(\varnothing) = 0$

Given a total order π of Ag and an agent $a \in Ag$:

$$\pi_{\geqslant a} = \{i \in Ag \mid \pi(i) \geqslant \pi(a)\}$$

$$\underbrace{val(\pi_{\geqslant a}) - val(\pi_{>a})}_{\text{contribution of agent } a \text{ to the value of coalition } \pi_{\geqslant a}}$$

Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $Ag = \{1, \ldots, n\}$,
- a payoff function $val : 2^{Ag} \to \mathbb{R}$ s.t. $val(\varnothing) = 0$

Given a total order π of Ag and an agent $a \in Ag$:

$$\pi_{\geqslant a} = \{i \in Ag \mid \pi(i) \geqslant \pi(a)\}$$

Shapley value: $$Sh(a) = \underbrace{\frac{1}{n!} \sum_{\pi \in \Pi_n} \Big(\underbrace{val(\pi_{\geqslant a}) - val(\pi_{>a})}_{\text{contribution of agent } a \text{ to the value of coalition } \pi_{\geqslant a}} \Big)}_{\text{“average contribution of agent } a\text{”}}$$

Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $Ag = \{1, \ldots, n\}$,
- a payoff function $val : 2^{Ag} \to \mathbb{R}$ s.t. $val(\varnothing) = 0$

Given a total order π of Ag and an agent $a \in Ag$:

$$\pi_{\geqslant a} = \{i \in Ag \mid \pi(i) \geqslant \pi(a)\}$$

Shapley value: $Sh(a) = \frac{1}{n!} \sum_{\pi \in \Pi_n} \left(val(\pi_{\geqslant a}) - val(\pi_{>a})\right)$

$$= \sum_{\substack{C \subseteq Ag \\ a \notin C}} \frac{|C|!(n-|C|-1)!}{n!} \left(val(C \cup \{a\}) - val(C)\right)$$

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

[Mascle/Baier/Funke/Jantsch/Kiefer, LICS'21]

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Goal: define a measure for the *impact of the states* $s \in S$ on the truth value of ϕ in terms of their nondeterministic choices.

[Mascle/Baier/Funke/Jantsch/Kiefer, LICS'21]

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Goal: define a measure for the *impact of the states* $s \in S$ on the truth value of ϕ in terms of their nondeterministic choices.

Game-based view:

- states may build coalitions that attempt to enforce ϕ
  no matter how the other states resolve their nondeterministic choices
- importance value of a state = Shapley value
  when the payoff is 1 for any coalition that can enforce ϕ and 0 otherwise

[Mascle/Baier/Funke/Jantsch/Kiefer, LICS'21]

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Let $C \subseteq S$ ... a coalition of states

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Let $C \subseteq S$ and $\mathcal{M}_C$ as before with objective ϕ for C

two-player turn-based game $\mathcal{M}_C$:

- arena: state space, initial state and transitions of $\mathcal{M}$
- player 1 controls all states in C (objective ϕ)
- player 2 controls all states in $\overline{C} = S \setminus C$ (objective $\neg\phi$)

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Let $C \subseteq S$ and $\mathcal{M}_C$ as before with objective ϕ for C

Payoff value of coalition C:

$$val_\phi(C) = \begin{cases} \mathbf{1} & : \text{ if } C \text{ has a winning strategy in } \mathcal{M}_C \text{ for } \phi \\ \mathbf{0} & : \text{ otherwise} \end{cases}$$

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Let $C \subseteq S$ and $\mathcal{M}_C$ as before with objective ϕ for C

Payoff value of coalition C:

$$val_\phi(C) = \begin{cases} \mathbf{1} : \text{ if } C \text{ has a winning strategy in } \mathcal{M}_C \text{ for } \phi \\ \mathbf{0} : \text{ otherwise} \end{cases}$$

Importance value of state s = Shapley value of s

in the simple cooperative game with agent set $Ag = S$ and payoff function val_ϕ

Importance values for path properties in TS

Given: a transition system $\mathcal{M}$ with state space S and initial state s_0 and a path property ϕ (e.g. LTL formula).

Let $C \subseteq S$ and $\mathcal{M}_C$ as before with objective ϕ for C

Payoff value of coalition C:

$$val_\phi(C) = \begin{cases} \mathbf{1} : \text{ if } C \text{ has a winning strategy in } \mathcal{M}_C \text{ for } \phi \\ \mathbf{0} : \text{ otherwise} \end{cases}$$

Importance value of state s = Shapley value of s

in the simple cooperative game with agent set $Ag = S$ and payoff function val_ϕ

0/1-values and monotonicity, i.e., if $C \subseteq D$ then $val_\phi(C) \leqslant val_\phi(D)$

Importance values: properties

Importance value of state s = Shapley value of s $\qquad n = |S|$

$$\mathcal{I}_\phi(s) \;=\; \sum_{\substack{C \subseteq S \\ s \notin C}} \frac{|C|!(n-|C|-1)!}{n!} \underbrace{\left(val_\phi(C \cup \{s\}) - val_\phi(C) \right)}_{0 \text{ or } 1}$$

Importance values: properties

Importance value of state s = Shapley value of s $\qquad n = |S|$

$$
\begin{aligned}
\mathcal{I}_\phi(s) &= \sum_{\substack{C \subseteq S \\ s \notin C}} \frac{|C|!(n-|C|-1)!}{n!} \big(\underbrace{val_\phi(C \cup \{s\}) - val_\phi(C)}_{0 \text{ or } 1}\big) \\
&= \sum_{\substack{(C,s) \\ \text{switching}}} \frac{|C|!(n-|C|-1)!}{n!}
\end{aligned}
$$

where (C, s) is switching iff $val_\phi(C \cup \{s\}) = 1$ and $val_\phi(C) = 0$

Importance values: properties

Importance value of state s = Shapley value of s $\qquad n = |S|$

$$\mathcal{I}_\phi(s) = \sum_{\substack{C \subseteq S \\ s \notin C}} \frac{|C|!(n-|C|-1)!}{n!}\left(val_\phi(C \cup \{s\}) - val_\phi(C)\right)$$

$$= \sum_{\substack{(C,s) \\ \text{switching}}} \frac{|C|!(n-|C|-1)!}{n!}$$

where (C, s) is switching iff $val_\phi(C \cup \{s\}) = 1$ and $val_\phi(C) = 0$

$\mathcal{I}_\phi(s) > 0$ iff s is relevant, i.e., there is a switching pair (C, s)

Importance values: properties

Importance value of state s = Shapley value of s $\quad n = |S|$

$$\mathcal{I}_\phi(s) = \sum_{\substack{C \subseteq S \\ s \notin C}} \frac{|C|!(n-|C|-1)!}{n!}\left(val_\phi(C\cup\{s\}) - val_\phi(C)\right)$$

$$= \sum_{\substack{(C,s) \\ \text{switching}}} \frac{|C|!(n-|C|-1)!}{n!} = \sum_{\substack{(C,s) \\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \quad \text{where } r = |R|$$

where (C, s) is switching iff $val_\phi(C\cup\{s\}) = 1$ and $val_\phi(C) = 0$

$\mathcal{I}_\phi(s) > 0$ iff s is relevant, i.e., there is a switching pair (C, s)

A switching pair (C, s) is relevant iff $C \subseteq R$ = set of relevant states

Importance values: properties

Importance value of state s = Shapley value of s

$$\mathcal{I}_{\phi}(s) \;=\; \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\text{ relevant states}$$

Zero-sum property of the game structure $\mathcal{M}_C$ yields:

$$val_{\phi}(C) \;=\; 1 - val_{\neg\phi}(\overline{C})$$

Importance values: properties

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) \;=\; \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\text{ relevant states}$$

Zero-sum property of the game structure $\mathcal{M}_C$ yields:

$$val_\phi(C) \;=\; 1 - val_{\neg\phi}(\overline{C})$$

$$(C, s) \text{ relevant for } \phi \quad \text{iff} \quad ((\overline{C} \cap R) \setminus \{s\}, s) \text{ relevant for } \neg\phi$$

Importance values: properties

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) \;=\; \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\text{ relevant states}$$

Zero-sum property of the game structure $\mathcal{M}_C$ yields:

$$val_\phi(C) \;=\; 1 - val_{\neg\phi}(\overline{C})$$

$$(C,s) \text{ relevant for } \phi \quad \text{iff} \quad (\underbrace{(\overline{C} \cap R) \setminus \{s\}}_{D}, s) \text{ relevant for } \neg\phi$$

$$|D| = r-|C|-1 \quad \text{and} \quad \frac{|C|!(r-|C|-1)!}{r!} = \frac{|D|!(r-|D|-1)!}{r!}$$

Importance values: properties

Importance value of state s = Shapley value of s

$$\mathcal{I}_{\phi}(s) \;=\; \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\ \text{relevant states}$$

Zero-sum property of the game structure $\mathcal{M}_C$ yields:

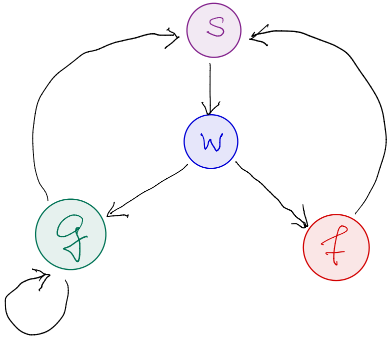
$$val_{\phi}(C) \;=\; 1 - val_{\neg\phi}(\overline{C})$$

$$(C,s) \text{ relevant for } \phi \quad \text{iff} \quad ((\overline{C} \cap R) \setminus \{s\}, s) \text{ relevant for } \neg\phi$$

Hence: $\quad \mathcal{I}_{\phi}(s) \;=\; \mathcal{I}_{\neg\phi}(s)$

"importance of states on the truth value (satisfaction or violation) of ϕ"

Importance values: example

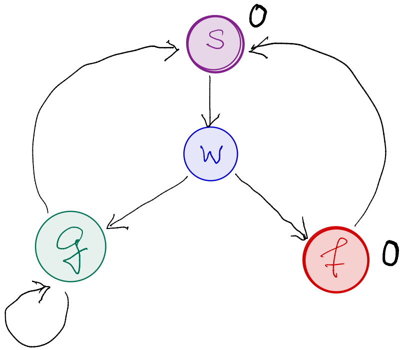


$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$ where r = # relevant states

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

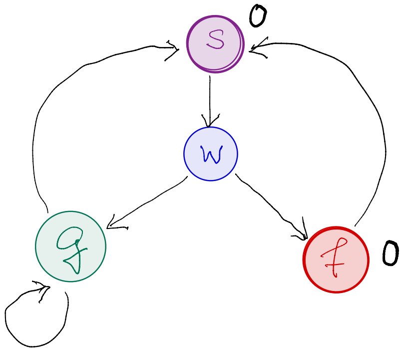
deterministic states are irrelevant
(importance value 0)

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$

where r = # relevant states

Importance values: example



$\phi = \Box\Diamond s \;\wedge\; \Diamond\Box\neg f$

deterministic states are irrelevant
(importance value 0)

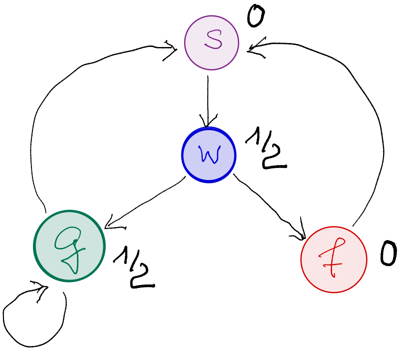
two relevant pairs: $(\{w\}, g)$, $(\{g\}, w)$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s) \\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$

where r = # relevant states

Importance values: example



$\phi = \Box\Diamond s \,\wedge\, \Diamond\Box\neg f$

deterministic states are irrelevant
(importance value 0)

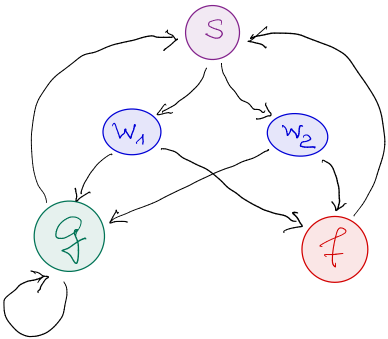
two relevant pairs: $(\{w\}, g)$, $(\{g\}, w)$

$$\mathcal{I}_\phi(w) = \mathcal{I}_\phi(g) = \frac{1!(2-1-1)!}{2!} = \frac{1!0!}{2!} = \frac{1}{2}$$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$ where r = # relevant states

Importance values: example

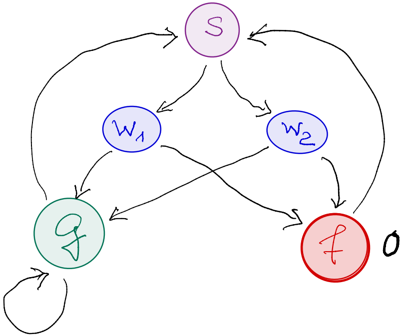


$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\text{ relevant states}$$

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

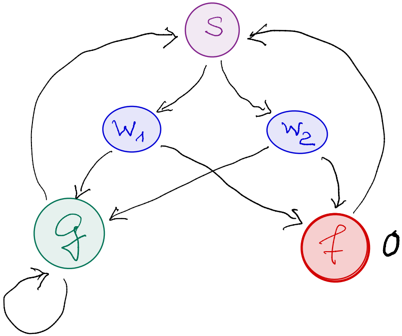
state f is irrelevant

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$

where r = # relevant states

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

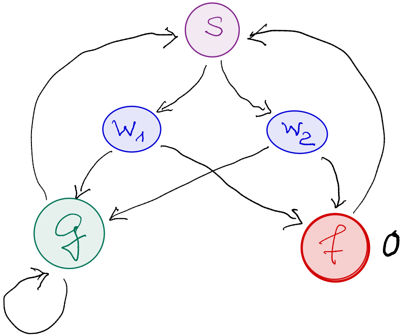
state f is irrelevant

C has a winning strategy iff
$g \in C$ and $\left|C \cap \{w_1, w_2, s\}\right| \geqslant 2$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s)\\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!} \qquad \text{where } r = \#\text{ relevant states}$$

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

state f is irrelevant

C has a winning strategy iff
$g \in C$ and $\left|C \cap \{w_1, w_2, s\}\right| \geqslant 2$

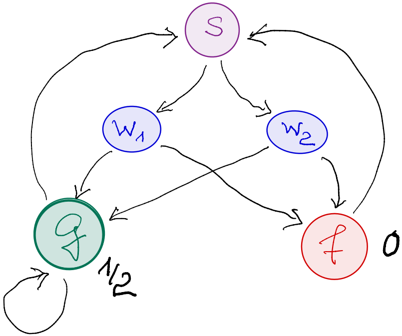
In particular: $r = 4$

Importance value of state s = Shapley value of s

$$\mathcal{I}_\phi(s) = \sum_{\substack{(C,s) \\ \text{relevant}}} \frac{|C|!(r-|C|-1)!}{r!}$$

where $r = \#$ relevant states

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

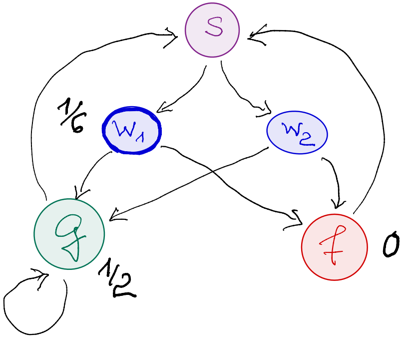
state f is irrelevant

C has a winning strategy iff
$g \in C$ and $\left|C \cap \{w_1, w_2, s\}\right| \geqslant 2$

In particular: $r = 4$

4 relevant pairs for g and $\mathcal{I}_\phi(g) = 3 \cdot \frac{2!(4-2-1)!}{4!} + \frac{3!(4-3-1)!}{4!} = \frac{1}{2}$

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

state f is irrelevant

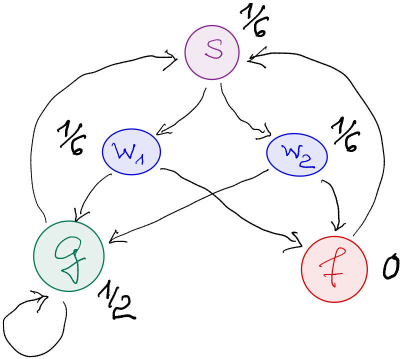
C has a winning strategy iff
$g \in C$ and $\left|C \cap \{w_1, w_2, s\}\right| \geqslant 2$

In particular: $r = 4$

4 relevant pairs for g and $\mathcal{I}_\phi(g) = 3 \cdot \frac{2!(4-2-1)!}{4!} + \frac{3!(4-3-1)!}{4!} = \frac{1}{2}$

2 relevant pairs for w_1 and $\mathcal{I}_\phi(w_1) = 2 \cdot \frac{2!(4-2-1)!}{4!} = \frac{1}{6}$

Importance values: example



$\phi = \Box\Diamond s \wedge \Diamond\Box\neg f$

state f is irrelevant

C has a winning strategy iff
$g \in C$ and $\big|C \cap \{w_1, w_2, s\}\big| \geqslant 2$

In particular: $r = 4$

4 relevant pairs for g and $\mathcal{I}_\phi(g) = 3 \cdot \frac{2!(4-2-1)!}{4!} + \frac{3!(4-3-1)!}{4!} = \frac{1}{2}$

2 relevant pairs for w_1 and $\mathcal{I}_\phi(w_1) = 2 \cdot \frac{2!(4-2-1)!}{4!} = \frac{1}{6}$

Importance values: algorithmic problems

For transition system $\mathcal{M}$ with state space S and path property ϕ.

Value problem:

given $C \subseteq S$, check whether $val_\phi(C) = 1$

Usefulness problem:

given state s, decide whether $\mathcal{I}_\phi(s) > 0$

Importance problem:

given state s, compute $n!\mathcal{I}_\phi(s)$

Importance values: algorithmic problems

For transition system $\mathcal{M}$ with state space S and path property ϕ.

Value problem: ... standard game solving

given $C \subseteq S$, check whether $val_{\phi}(C) = 1$

Usefulness problem:

given state s, decide whether $\mathcal{I}_{\phi}(s) > 0$

Importance problem:

given state s, compute $n!\,\mathcal{I}_{\phi}(s)$

Importance values: algorithmic problems

For transition system $\mathcal{M}$ with state space S and path property ϕ.

Value problem: ... standard game solving

given $C \subseteq S$, check whether $val_\phi(C) = 1$

Usefulness problem:

given state s, decide whether $\mathcal{I}_\phi(s) > 0$

Importance problem:

given state s, compute $n!\mathcal{I}_\phi(s)$

Solving the usefulness and importance problems, via standard game solving algorithms + guessing relevant pairs.

Importance values: complexity results

	Büchi	Rabin	Streett	Parity	LTL
Value problem	P	NP	coNP	$\in \mathrm{NP} \cap \mathrm{coNP}$	$\mathrm{2EXP}$
Usefulness problem	NP	Σ_2^P	Σ_2^P	NP	$\mathrm{2EXP}$
Importance problem	$\#\mathrm{P}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}$	$\mathrm{2EXP}$

Importance values: complexity results

	Büchi	Rabin	Streett	Parity	LTL
Value problem	P	NP	coNP	$\in \mathrm{NP} \cap \mathrm{coNP}$	$\mathrm{2EXP}$
Usefulness problem	NP	Σ_2^P	Σ_2^P	NP	$\mathrm{2EXP}$
Importance problem	$\#\mathrm{P}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}$	$\mathrm{2EXP}$

Value problem: classical results for games

Importance values: complexity results

	Büchi	Rabin	Streett	Parity	LTL
Value problem	P	NP	coNP	$\in \mathrm{NP} \cap \mathrm{coNP}$	$\mathrm{2EXP}$
Usefulness problem	NP	Σ_2^P	Σ_2^P	NP	$\mathrm{2EXP}$
Importance problem	$\#\mathrm{P}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}$	$\mathrm{2EXP}$

NP-completeness of the usefulness problem for Büchi conditions

- upper bound via guess-&-check method

  nondeterministically guess a set C and check whether (C, s) is relevant (with poly-time algorithm for Büchi games)

- NP-hardness via reduction from 3SAT

Importance values: complexity results

	Büchi	Rabin	Streett	Parity	LTL
Value problem	P	NP	coNP	$\in \mathrm{NP} \cap \mathrm{coNP}$	$2\mathrm{EXP}$
Usefulness problem	NP	Σ_2^P	Σ_2^P	NP	$2\mathrm{EXP}$
Importance problem	$\#\mathrm{P}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}^{\mathrm{NP}}$	$\#\mathrm{P}$	$2\mathrm{EXP}$

Σ_2^p-completeness of the usefulness problem for Rabin conditions

- upper bound via guess-&-check method

  nondeterministically guess a set C and check whether (C, s) is relevant (with NP-oracle for Rabin games)

- Σ_2^p-hardness via reduction from dual of $\forall\exists$3SAT

Break

Outline

- Introduction
- Necessary and sufficient causes
- Counterfactuality and responsibility in verification
- Probabilistic causality in Markovian models
- Conclusions

Probabilistic causality

Probabilistic causality

... extensively studied in philosophy

Reichenbach (1956)
Suppes (1970)
and many more

PDF version of the entry
Hans Reichenbach
https://plato.stanford.edu/archives/spr2022/entries/reichenbach/
from the Spring 2022 Edition of the

STANFORD ENCYCLOPEDIA OF PHILOSOPHY

Library of Congress ISSN: 1095-5054

Photo from the Hans Reichenbach Collection, reproduced by permission of the University of Pittsburgh. All rights reserved.

Hans Reichenbach

First published Sun Aug 24, 2008; substantive revision Tue Mar 23, 2021

Described as perhaps "the greatest empiricist of the 20th century" (Salmon, 1977a), the work of Hans Reichenbach (1891–1953) provides one of the main statements of empiricist philosophy in the 20th century. Provoked by the conflict between (neo-) Kantian a priorism and Einstein's relativity of space and time, Reichenbach developed a scientifically inspired philosophy and an uncompromisingly empiricist epistemology. He was literate in the physical science of his time, and acquainted with many of its most eminent practitioners. Criticism and justification of scientific methodology formed the core of almost all his philosophical efforts, which he promoted in a crescendo of books, in the journal *Erkenntnis*, which he founded and edited with Rudolf Carnap, and within a group of philosophers, mathematicians and scientists he led in Berlin. His commitment to objectivity and realism in science together with his probabilistic justification of belief in scientific results carried philosophical and technical difficulties that shaped much of the subsequent debate in philosophy of science. Reichenbach's contributions cover large swathes of formal philosophy, especially in philosophy of physics, logic, induction and the foundations of probability, and his later work encompassed

1

Probabilistic causality

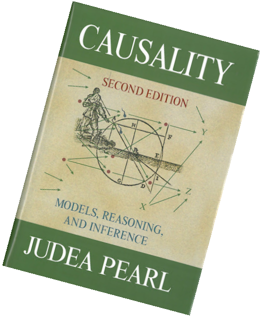
... extensively studied in philosophy, but also in AI

Reichenbach (1956)
Suppes (1970)
and many more

Judea Pearl

Turing Award
Winner 2011

taken from Judea Pearl's homepage
UCLA Cognitive Systems Laboratory



Probabilistic causality

... extensively studied in philosophy, but also in AI

Two main principles:

> *Temporal condition:*
>
> *Probability-raising condition:*

Probabilistic causality

... extensively studied in philosophy, but also in AI

Two main principles:

Temporal condition:

Causes occur before their effects.

Probability-raising condition:

Probabilistic causality

… extensively studied in philosophy, but also in AI

Two main principles:

Temporal condition:

Causes occur before their effects.

Probability-raising condition:

$$\Pr(\, \mathit{effect} \mid \mathit{cause} \,) > \Pr(\, \mathit{effect} \mid \neg\, \mathit{cause} \,)$$

Probabilistic causality

... extensively studied in philosophy, but also in AI

Two main principles:

Temporal condition:

Causes occur before their effects.

Probability-raising condition:

$$\mathrm{Pr}(\ \textit{effect} \mid \textit{cause}\ ) > \mathrm{Pr}(\ \textit{effect} \mid \neg\,\textit{cause}\ )$$

equivalently: $\mathrm{Pr}(\ \textit{effect} \mid \textit{cause}\ ) > \mathrm{Pr}(\ \textit{effect}\ )$

Probabilistic causality

... extensively studied in philosophy, but also in AI

Two main principles:

Temporal condition:

Causes occur before their effects.

Probability-raising condition:

$$\Pr(\ \mathit{effect} \mid \mathit{cause}\ ) \ > \ \Pr(\ \mathit{effect} \mid \neg\, \mathit{cause}\ )$$

probabilistic form of counterfactuality:
"effects are less likely if their causes do not occur"

Probabilistic causality in operational models

Probabilistic causality in operational models

Only very few research so far:

- formalization for sets of states by PCTL-constraints in Markov chains [Kleinberg, PhD thesis 2010]

PCTL: probabilistic computation tree logic

Probabilistic causality in operational models

Only very few research so far:

- formalization for sets of states by PCTL-constraints in Markov chains [Kleinberg, PhD thesis 2010]
- formalization as probabilistic hyperproperties

  in Markov chains [Ábrahám/Bonakdarpour, QEST'18]

  in Markov decision processes [Dimitrova/Finkbeiner/Torfah, ATVA'20]

PCTL: probabilistic computation tree logic

Probabilistic causality in operational models

Only very few research so far:

- formalization for sets of states by PCTL-constraints in Markov chains [Kleinberg, PhD thesis 2010]
- formalization as probabilistic hyperproperties

  in Markov chains [Ábrahám/Bonakdarpour, QEST'18]

  in Markov decision processes [Dimitrova/Finkbeiner/Torfah, ATVA'20]
- cause-effect relations for regular causes and ω-regular effects in Markov chains [B./Funke/Jantsch/Piribauer/Ziemek, ATVA'21]

PCTL: probabilistic computation tree logic

Probabilistic causality in operational models

Only very few research so far:

- formalization for sets of states by PCTL-constraints in Markov chains [Kleinberg, PhD thesis 2010]
- formalization as probabilistic hyperproperties
  - in Markov chains [Ábrahám/Bonakdarpour, QEST'18]
  - in Markov decision processes [Dimitrova/Finkbeiner/Torfah, ATVA'20]
- cause-effect relations for regular causes and ω-regular effects in Markov chains [B./Funke/Jantsch/Piribauer/Ziemek, ATVA'21]
- cause-effect relations for sets of states in Markov decision processes [B./Funke/Piribauer/Ziemek, FoSSaCS'22]

Probabilistic causality in operational models

Only very few research so far:

- formalization for sets of states by PCTL-constraints in Markov chains [Kleinberg, PhD thesis 2010]
- formalization as probabilistic hyperproperties

  in Markov chains [Ábrahám/Bonakdarpour, QEST'18]

  in Markov decision processes [Dimitrova/Finkbeiner/Torfah, ATVA'20]
- cause-effect relations for regular causes and ω-regular effects in Markov chains [B./Funke/Jantsch/Piribauer/Ziemek, ATVA'21]
- cause-effect relations for sets of states in Markov decision processes [B./Funke/Piribauer/Ziemek, FoSSaCS'22]

Probabilistic causality in Markov chains

Probabilistic causality in Markov chains

In what follows: $\mathcal{M}$ is a (discrete-time) Markov chain with

- finite state space S
- initial distribution $\iota : S \to [0, 1]$ such that every state in S is accessible from at least one initial state (i.e., a state s with $\iota(s) > 0$)
- a fixed nonempty set E of effect states

Probabilistic causality in Markov chains

In what follows: $\mathcal{M}$ is a (discrete-time) Markov chain with

- finite state space S
- initial distribution $\iota : S \to [0,1]$ such that every state in S is accessible from at least one initial state (i.e., a state s with $\iota(s) > 0$)
- a fixed nonempty set E of effect states

W.l.o.g. all E-states are terminal (i.e., do not have outgoing transitions).

Probabilistic causality in Markov chains

In what follows: $\mathcal{M}$ is a (discrete-time) Markov chain with

- finite state space S
- initial distribution $\iota : S \to [0,1]$ such that every state in S is accessible from at least one initial state (i.e., a state s with $\iota(s) > 0$)
- a fixed nonempty set E of effect states

W.l.o.g. all E-states are terminal (i.e., do not have outgoing transitions).

$\Pr_{\mathcal{M}}(\Diamond E)$	effect probability in $\mathcal{M}$
$\Pr_s(\Diamond E)$	effect probability from state s

Probabilistic causality in Markov chains

In what follows: $\mathcal{M}$ is a (discrete-time) Markov chain with

- finite state space S
- initial distribution $\iota : S \to [0, 1]$ such that every state in S is accessible from at least one initial state (i.e., a state s with $\iota(s) > 0$)
- a fixed nonempty set E of effect states

W.l.o.g. all E-states are terminal (i.e., do not have outgoing transitions).

$\Pr_{\mathcal{M}}(\Diamond E)$	effect probability in $\mathcal{M}$	$= \sum\limits_{s \in S} \iota(s) \cdot \Pr_s(\Diamond E)$
$\Pr_s(\Diamond E)$	effect probability from state s	

PCTL-characterization of causality in MC

[Kleinberg, PhD thesis 2010]

PCTL: probabilistic computation logic

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in \,]0,1]$ s.t.

$$\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E) \quad \text{and} \quad \mathcal{M} \models \forall\Box\big(C \to \mathbb{P}_{\geqslant p}(\Diamond E)\big)$$

[Kleinberg, PhD thesis 2010]

PCTL: probabilistic computation logic

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in]0,1]$ s.t.

$$\underbrace{\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E)}_{\Pr_{\mathcal{M}}(\Diamond E) < p} \quad \text{and} \quad \mathcal{M} \models \forall\Box(C \to \mathbb{P}_{\geqslant p}(\Diamond E))$$

[Kleinberg, PhD thesis 2010]

PCTL: probabilistic computation logic

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in]0,1]$ s.t.

$$\underbrace{\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E)}_{\Pr_{\mathcal{M}}(\Diamond E) < p} \quad \text{and} \quad \underbrace{\mathcal{M} \models \forall\Box\big(C \to \mathbb{P}_{\geqslant p}(\Diamond E)\big)}_{\substack{\Pr_s(\Diamond E) \geqslant p \\ \text{for all } s \in C}}$$

[Kleinberg, PhD thesis 2010]

PCTL: probabilistic computation logic

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in]0,1]$ s.t.

$$\underbrace{\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E)}_{\Pr_{\mathcal{M}}(\Diamond E) < p} \quad \text{and} \quad \underbrace{\mathcal{M} \models \forall\Box\big(C \rightarrow \mathbb{P}_{\geqslant p}(\Diamond E)\big)}_{\substack{\Pr_s(\Diamond E) \geqslant p \\ \text{for all } s \in C}}$$

Thus:

C cause for E iff $\Pr_{\mathcal{M}}(\Diamond E) < \Pr_s(\Diamond E)$ for all $s \in C$

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in]0,1]$ s.t.

$$\underbrace{\mathcal{M} \models \mathbb{P}_{<p}(\lozenge E)}_{\Pr_{\mathcal{M}}(\lozenge E) < p} \quad \text{and} \quad \underbrace{\mathcal{M} \models \forall\Box(C \rightarrow \mathbb{P}_{\geqslant p}(\lozenge E))}_{\substack{\Pr_s(\lozenge E) \geqslant p \\ \text{for all } s \in C}}$$

Thus:

$$\begin{array}{lll} C \text{ cause for } E & \text{iff} \quad \Pr_{\mathcal{M}}(\lozenge E) \ < \ \Pr_s(\lozenge E) & \text{for all } s \in C \\ & \text{iff} \quad \Pr_{\mathcal{M}}(\lozenge E) < \Pr_{\mathcal{M}}(\lozenge E | \lozenge s) & \text{for all } s \in C \end{array}$$

PCTL-characterization of causality in MC

Let C a set of states with $C \cap E = \varnothing$.

C is called a *(prima facie) cause* for E if there exists $p \in \,]0,1]$ s.t.

$$\underbrace{\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E)}_{\Pr_{\mathcal{M}}(\Diamond E) < p} \quad \text{and} \quad \underbrace{\mathcal{M} \models \forall\Box(C \rightarrow \mathbb{P}_{\geqslant p}(\Diamond E))}_{\substack{\Pr_s(\Diamond E) \geqslant p \\ \text{for all } s \in C}}$$

Thus:

C cause for E iff $\Pr_{\mathcal{M}}(\Diamond E) < \Pr_s(\Diamond E)$ for all $s \in C$

iff $\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\Diamond E | \Diamond s)$ for all $s \in C$

strict probability-raising condition

(elementwise for all C-states)

Strict/global probability-raising causes in MC

Let C a set of states with $C \cap E = \varnothing$.

- C is a *strict probability-raising (SPR) cause* for E iff

$$\mathrm{Pr}_{\mathcal{M}}(\Diamond E) \; < \; \underbrace{\mathrm{Pr}_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond s\,)}_{\mathrm{Pr}_s(\Diamond E)} \;\; \text{for all } s \in C$$

Strict/global probability-raising causes in MC

Let C a set of states with $C \cap E = \varnothing$.

- C is a *strict probability-raising (SPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) \; < \; \Pr_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond s\,) \quad \text{for all } s \in C$$

- C is a *global probability-raising (GPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) \; < \; \underbrace{\Pr_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond C\,)}_{\substack{\text{conditional probability} \\ \frac{\sum_{s \in C} \Pr_{\mathcal{M}}((\neg C) \, \mathsf{U} \, s) \cdot \Pr_s(\Diamond E)}{\Pr_{\mathcal{M}}(\Diamond C)}}}$$

Strict/global probability-raising causes in MC

Let C a set of states with $C \cap E = \varnothing$.

- C is a *strict probability-raising (SPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) \;<\; \Pr_{\mathcal{M}}(\Diamond E \,|\, \Diamond s) \quad \text{for all } s \in C$$

- C is a *global probability-raising (GPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) \;<\; \Pr_{\mathcal{M}}(\Diamond E \,|\, \Diamond C)$$

plus $\underbrace{\text{some minimality constraint}}$ (omitted here)

"no C-state is fully covered by other C-states"

i.e., for each state $s \in C$ there is a path π in $\mathcal{M}$ with $\pi \models (\neg C) \, \mathrm{U} \, s$.

Strict/global probability-raising causes in MC

Let C a set of states with $C \cap E = \varnothing$.

- C is a *strict probability-raising (SPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\Diamond E \mid \Diamond s) \quad \text{for all } s \in C$$

- C is a *global probability-raising (GPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\Diamond E \mid \Diamond C)$$

plus some minimality constraint (omitted here)

- Each SPR cause is a GPR cause.

Strict/global probability-raising causes in MC

Let C a set of states with $C \cap E = \varnothing$.

- C is a *strict probability-raising (SPR) cause* for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) \; < \; \Pr_{\mathcal{M}}(\, \Diamond E \,|\, \Diamond s \,) \quad \text{for all } s \in C$$

- C is a *global probability-raising (GPR) cause* for E iff

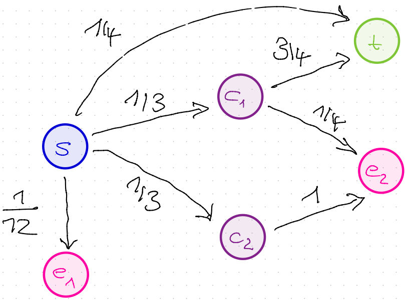
$$\Pr_{\mathcal{M}}(\Diamond E) \; < \; \Pr_{\mathcal{M}}(\, \Diamond E \,|\, \Diamond C \,)$$

plus some minimality constraint (omitted here)

- Each SPR cause is a GPR cause.
- If C is a singleton then:

C is a SPR cause iff C is a GPR cause

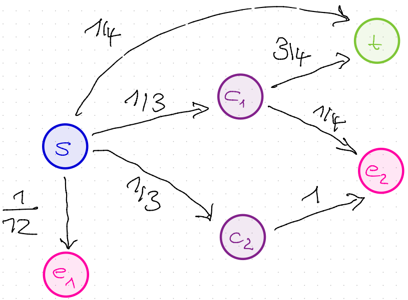
Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s

effect set $E = \{e_1, e_2\}$

Example: PR cause in MC

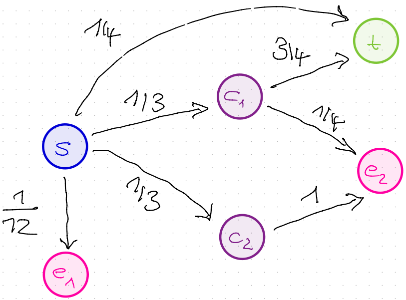


MC $\mathcal{M}$ with unique initial state s

effect set $E = \{e_1, e_2\}$

$$\Pr_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{3} + \frac{1}{3}\cdot\frac{1}{4} + \frac{1}{12} \;=\; \frac{1}{2}$$

Example: PR cause in MC



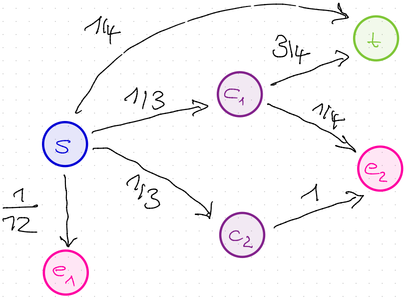
MC $\mathcal{M}$ with unique initial state s

effect set $E = \{e_1, e_2\}$

$$\Pr_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{3} + \frac{1}{3}\cdot\frac{1}{4} + \frac{1}{12} \;=\; \frac{1}{2}$$

$C = \{c_1, c_2\}$

Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s

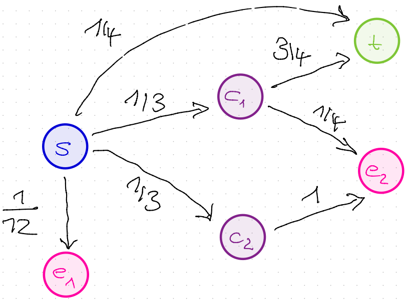
effect set $E = \{e_1, e_2\}$

$\Pr_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{12} \;=\; \frac{1}{2}$

$C = \{c_1, c_2\}$

- C is not an SPR cause as $\Pr_{c_1}(\Diamond E) = \frac{1}{4} < \frac{1}{2} \;=\; \Pr_{\mathcal{M}}(\Diamond E)$

Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s

effect set $E = \{e_1, e_2\}$

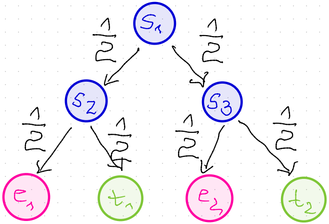
$$\Pr_{\mathcal{M}}(\Diamond E) \;=\; \tfrac{1}{3} + \tfrac{1}{3}\cdot\tfrac{1}{4} + \tfrac{1}{12} \;=\; \tfrac{1}{2}$$

$C = \{c_1, c_2\}$

- C is not an SPR cause as $\Pr_{c_1}(\Diamond E) = \frac{1}{4} < \frac{1}{2} \;=\; \Pr_{\mathcal{M}}(\Diamond E)$
- C is a GPR cause as

$$\Pr_{\mathcal{M}}(\Diamond E \mid \Diamond C) \;=\; \frac{\frac{1}{3}+\frac{1}{3}\cdot\frac{1}{4}}{\frac{1}{3}+\frac{1}{3}} \;=\; \frac{\frac{5}{12}}{\frac{2}{3}} \;=\; \frac{5}{8} \;>\; \frac{1}{2} \;=\; \Pr_{\mathcal{M}}(\Diamond E)$$

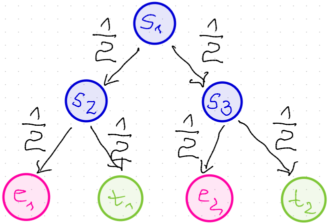
Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s_1

effect set $E = \{e_1, e_2\}$

Example: PR cause in MC



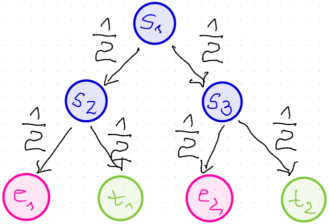
MC $\mathcal{M}$ with unique initial state s_1

effect set $E = \{e_1, e_2\}$

$$\Pr_{\mathcal{M}}(\Diamond E) = \Pr_s(\Diamond E) = \frac{1}{2}$$

for each state $s \in \{s_1, s_2, s_3\}$

Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s_1

effect set $E = \{e_1, e_2\}$

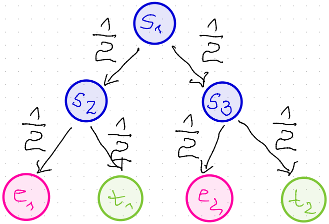
$\Pr_{\mathcal{M}}(\Diamond E) \;=\; \Pr_s(\Diamond E) \;=\; \frac{1}{2}$

for each state $s \in \{s_1, s_2, s_3\}$

There is no GPR cause as for any $C \subseteq \{s_1, s_2, s_3\}$:

$$\Pr_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond C\,) \;=\; \frac{1}{2} \;=\; \Pr_{\mathcal{M}}(\Diamond E)$$

Example: PR cause in MC



MC $\mathcal{M}$ with unique initial state s_1

effect set $E = \{e_1, e_2\}$

$\Pr_{\mathcal{M}}(\Diamond E) = \Pr_s(\Diamond E) = \frac{1}{2}$

for each state $s \in \{s_1, s_2, s_3\}$

There is no GPR cause as for any $C \subseteq \{s_1, s_2, s_3\}$:

$$\Pr_{\mathcal{M}}(\Diamond E \mid \Diamond C) = \frac{1}{2} = \Pr_{\mathcal{M}}(\Diamond E)$$

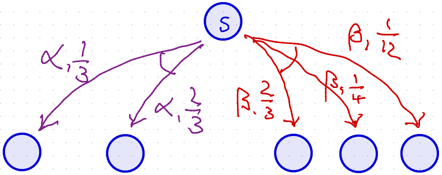
Well justified, as the events $\Diamond E$ and $\Diamond C$ are stochastically independent for any C.

Markov decision processes (MDP)

... extension of Markov chains by nondeterministic choices ...

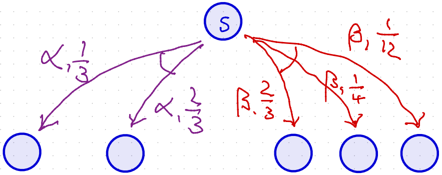
Markov decision processes (MDP)

- finite state space S with initial distribution $\iota : S \to [0,1]$
- finite set of action Act
- for each state $s \in S$:
  - $Act(s)$: set of enabled actions in state s
  - for each action $\alpha \in Act(s)$: distribution $P_{s,\alpha} : S \to [0,1]$ for the α-successors of s



Markov decision processes (MDP)

- finite state space S with initial distribution $\iota : S \to [0,1]$
- finite set of action Act
- for each state $s \in S$:
  - $Act(s)$: set of enabled actions in state s
  - for each action $\alpha \in Act(s)$: distribution $P_{s,\alpha} : S \to [0,1]$



Scheduler (a.k.a. policy, adversary, strategy): resolves the nondeterminism

- selects distributions over enabled actions (might be history-dependent)
- induced stochastic process is a Markov chain (tree-like, possibly infinite)

PR causes in MDPs

... generalize the definition of SPR and GPR causes for MDPs ...

PR causes in MDPs

... generalize the definition of SPR and GPR causes for MDPs ...

Assumptions: given an MDP $\mathcal{M}$ with state space S and:

- fixed effect set E consisting of terminal states
  (i.e., have no enabled action)

PR causes in MDPs

... generalize the definition of SPR and GPR causes for MDPs ...

Assumptions: given an MDP $\mathcal{M}$ with state space S and:

- fixed effect set E consisting of terminal states
  (i.e., have no enabled action)
- all states in S are reachable from at least one initial state

PR causes in MDPs

... generalize the definition of SPR and GPR causes for MDPs ...

Assumptions: given an MDP $\mathcal{M}$ with state space S and:

- fixed effect set E consisting of terminal states
  (i.e., have no enabled action)
- all states in S are reachable from at least one initial state
- all states in S from which E is not reachable are terminal

PR causes in MCs (repetition)

Let C a set of states with $C \cap E = \varnothing$. C is a

- SPR cause for E iff for all $s \in C$

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\,\Diamond E \mid \Diamond s\,)$$

- GPR cause for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\,\Diamond E \mid \Diamond C\,)$$

SPR: strict probability-raising
GPR: global probability-raising

PR causes in MCs (repetition)

Let C a set of states with $C \cap E = \varnothing$. C is a

- SPR cause for E iff for all $s \in C$

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\Diamond E \mid (\neg C) \, \mathsf{U} \, s)$$

- GPR cause for E iff

$$\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{\mathcal{M}}(\Diamond E \mid \Diamond C)$$

SPR: strict probability-raising
GPR: global probability-raising

PR causes in MDPs

Let C a set of states with $C \cap E = \varnothing$. C is a

- SPR cause for E iff for all $s \in C$ and all schedulers σ:

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) < \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \,|\, (\neg C) \,\mathsf{U}\, s\,)$$

- GPR cause for E iff for all schedulers σ:

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) < \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond C\,)$$

$$\Pr^{\sigma}_{\mathcal{M}}(\ldots) = \left\{ \begin{array}{l} \text{probability measure of the Markov chain} \\ \text{induced by scheduler } \sigma \end{array} \right.$$

PR causes in MDPs

Let C a set of states with $C \cap E = \varnothing$. C is a

- SPR cause for E iff for all $s \in C$ and all schedulers σ:

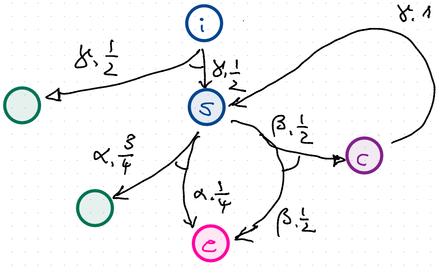
$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) < \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \,|\, (\neg C)\,\mathsf{U}\,s\,) \quad \text{if } \Pr^{\sigma}_{\mathcal{M}}((\neg C)\,\mathsf{U}\,s) > 0$$

- GPR cause for E iff for all schedulers σ:

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) < \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond C\,) \qquad \text{if } \Pr^{\sigma}_{\mathcal{M}}(\Diamond C) > 0$$

$$\Pr^{\sigma}_{\mathcal{M}}(\ldots) = \left\{ \begin{array}{l} \text{probability measure of the Markov chain} \\ \text{induced by scheduler } \sigma \end{array} \right.$$

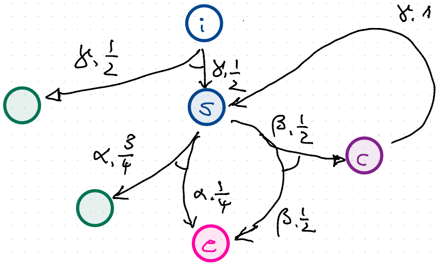
Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

Example: PR cause in MDP

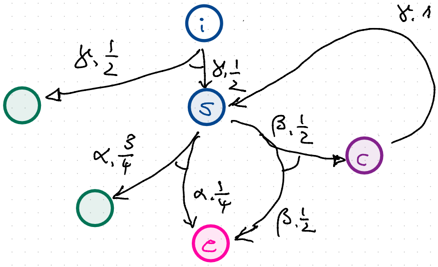


MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

Is $C = \{c\}$ a PR cause?

Example: PR cause in MDP



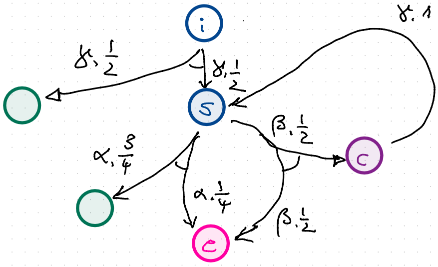
MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

Is $C = \{c\}$ a PR cause?

No

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

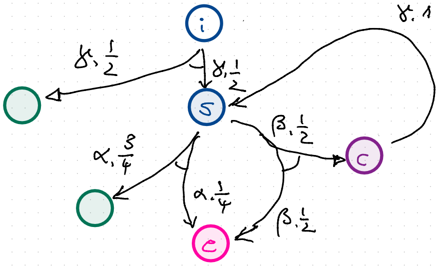
effect set $E = \{e\}$

Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules β for the first visit of s and α for the second visit of s.

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

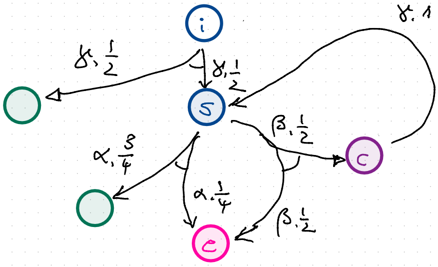
Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules β for the first visit of s and α for the second visit of s.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\cdot 1\cdot\frac{1}{4} \;=\; \frac{5}{16}$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

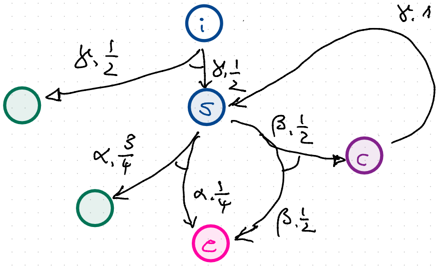
Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules β for the first visit of s and α for the second visit of s.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{4} = \frac{5}{16} > \frac{1}{4} = \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond c)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

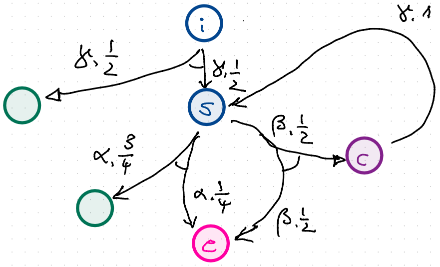
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider the scheduler σ that schedules β for the first visit of s and α for the second visit of s.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\cdot 1\cdot\frac{1}{4} = \frac{5}{16} > \frac{1}{4} = \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \,|\, \Diamond c\,)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

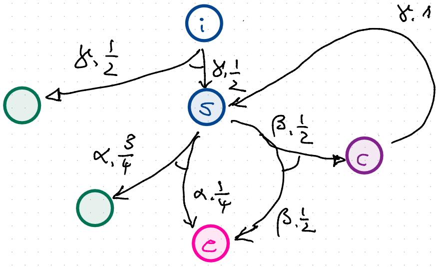
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1 - \lambda$.

MR = memoryless randomized

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

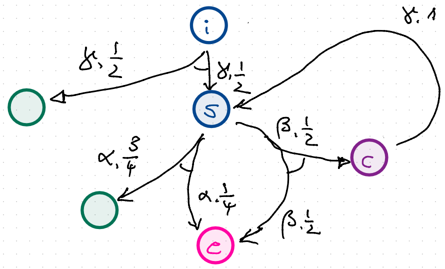
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1-\lambda$.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) = \frac{1}{2}\cdot\Pr^{\sigma}_{s}(\Diamond E)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

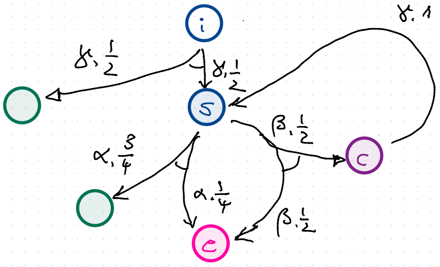
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1 - \lambda$.

$$\mathrm{Pr}^{\sigma}_{\mathcal{M}}(\Diamond E) = \tfrac{1}{2} \cdot \underbrace{\mathrm{Pr}^{\sigma}_{s}(\Diamond E)}_{\text{some positive value}} < \mathrm{Pr}^{\sigma}_{s}(\Diamond E)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

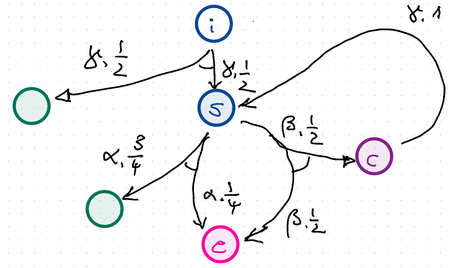
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1-\lambda$.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \underbrace{\Pr^{\sigma}_{s}(\Diamond E)}_{\text{some positive value}} < \Pr^{\sigma}_{s}(\Diamond E) = \Pr^{\sigma}_{c}(\Diamond E)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

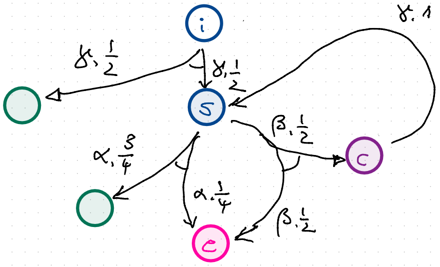
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1-\lambda$.

$$\mathrm{Pr}^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2}\cdot\underbrace{\mathrm{Pr}^\sigma_s(\Diamond E)}_{\text{some positive value}} < \mathrm{Pr}^\sigma_s(\Diamond E) = \mathrm{Pr}^\sigma_c(\Diamond E) = \mathrm{Pr}^\sigma_{\mathcal{M}}(\,\Diamond E \mid \Diamond c\,)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e\}$

Is $C = \{c\}$ a PR cause?

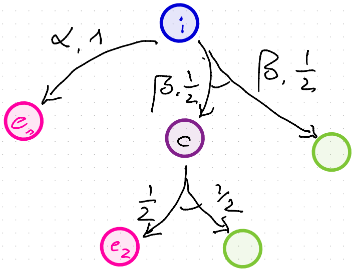
No, although PR condition holds for all memoryless schedulers

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1-\lambda$.

$$\Pr^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2}\cdot\Pr^\sigma_s(\Diamond E) < \Pr^\sigma_s(\Diamond E) = \Pr^\sigma_c(\Diamond E) = \Pr^\sigma_{\mathcal{M}}(\,\Diamond E \mid \Diamond c\,)$$

Consequence: Memory can be needed for refuting the PR condition!

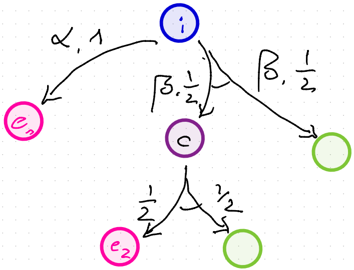
Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Example: PR cause in MDP

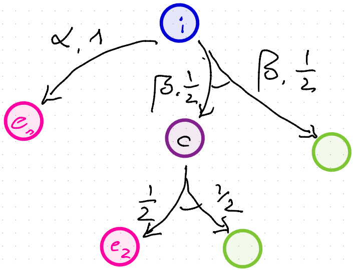


MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

Example: PR cause in MDP



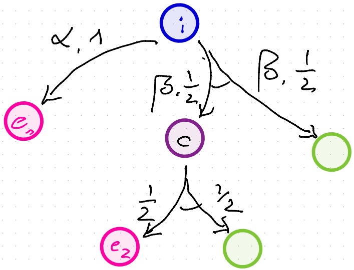
MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

No

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

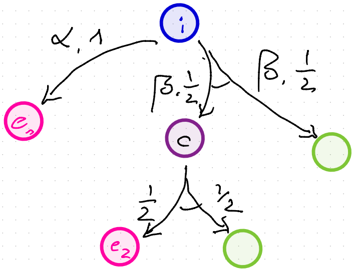
effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules α and β with probability $1/2$ in state i.

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

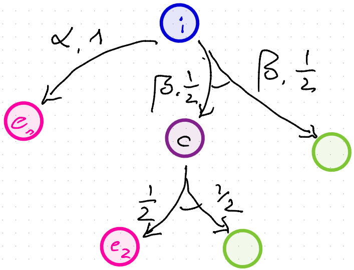
Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules α and β with probability $1/2$ in state i.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} \;=\; \frac{5}{8}$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

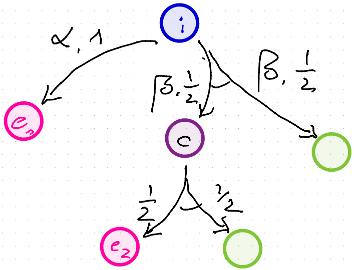
Is $C = \{c\}$ a PR cause?

No

Consider the scheduler σ that schedules α and β with probability $1/2$ in state i.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} \;=\; \frac{5}{8} \;>\; \frac{1}{2} \;=\; \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \,|\, \Diamond c)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

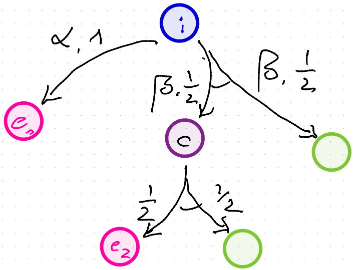
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all deterministic schedulers

Consider the scheduler σ that schedules α and β with probability $1/2$ in state i.

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8} > \frac{1}{2} = \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond c)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

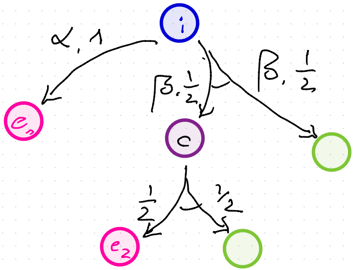
effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all deterministic schedulers

Consider the deterministic schedulers σ_α and σ_β that schedule α resp. β in state i.

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

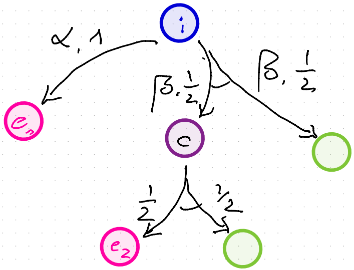
Is $C = \{c\}$ a PR cause?

No, although PR condition holds for all deterministic schedulers

Consider the deterministic schedulers σ_α and σ_β that schedule α resp. β in state i.

σ_α irrelevant for PR condition as state c is not reachable

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

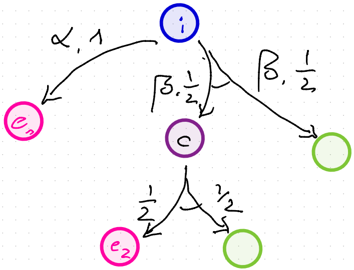
No, although PR condition holds for all deterministic schedulers

Consider the deterministic schedulers σ_α and σ_β that schedule α resp. β in state i.

σ_α irrelevant for PR condition as state c is not reachable

$$\Pr^{\sigma_\beta}_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} < \frac{1}{2} = \Pr^{\sigma_\beta}_{\mathcal{M}}(\Diamond E \mid \Diamond c)$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

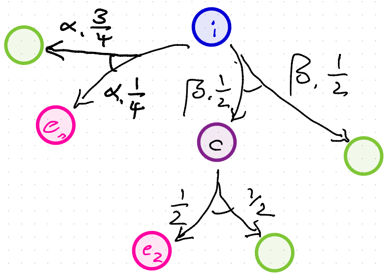
No, although PR condition holds for all deterministic schedulers

Consider the deterministic schedulers σ_α and σ_β that schedule α resp. β in state i.

⋮

Consequence: Randomization needed for refuting the PR condition!

Example: PR cause in MDP



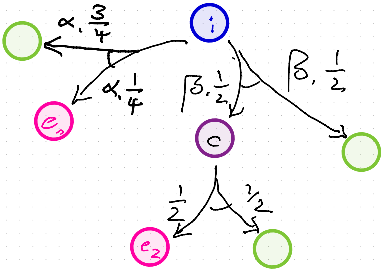
MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

Yes !!

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

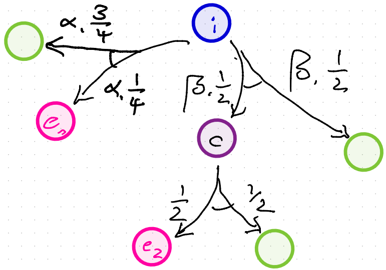
effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

Yes !!

Let σ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1-\lambda$.

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

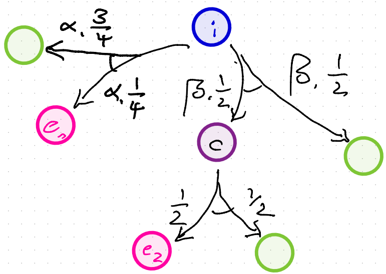
Is $C = \{c\}$ a PR cause?

Yes !!

Let σ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1 - \lambda$.

If $\lambda = 1$ then σ is irrelevant (as c is not reachable along σ-paths).

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

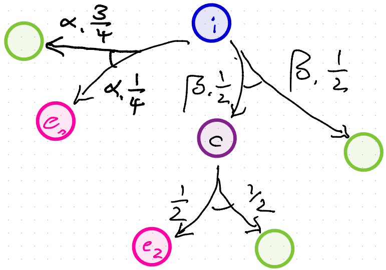
Yes !!

Let σ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1-\lambda$.

If $\lambda = 1$ then σ is irrelevant (as c is not reachable along σ-paths). Otherwise:

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{4}\cdot\lambda + \frac{1}{2}\cdot\frac{1}{2}\cdot(1-\lambda) \;=\; \frac{1}{4}$$

Example: PR cause in MDP



MDP $\mathcal{M}$ with unique initial state i

effect set $E = \{e_1, e_2\}$

Is $C = \{c\}$ a PR cause?

Yes !!

Let σ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1-\lambda$.

If $\lambda = 1$ then σ is irrelevant (as c is not reachable along σ-paths). Otherwise:

$$\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) \;=\; \frac{1}{4}\cdot\lambda + \frac{1}{2}\cdot\frac{1}{2}\cdot(1-\lambda) \;=\; \frac{1}{4} \;<\; \frac{1}{2} \;=\; \Pr^{\sigma}_{\mathcal{M}}(\,\Diamond E \mid \Diamond c\,)$$

Algorithmic problems

Algorithmic problems

Checking cause-effect relationships:

Finding good causes for given effects:

Algorithmic problems

Checking cause-effect relationships:

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}$, E, C, check whether

- C is an SPR cause for E
- C is a GPR cause for E

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}$, E, C, check whether

- C is an SPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

- C is a GPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}$, E, C, check whether

- C is an SPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: poly-time by statewise checking of the SPR condition

- C is a GPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}$, E, C, check whether

- C is an SPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: poly-time by statewise checking of the SPR condition

- C is a GPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: in PSPACE, using an encoding of the violation of the GPR condition in ETR (quadratic + linear constraints)

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components

  i.e., under all schedulers a terminal state will eventually be reached a.s.

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components
  i.e., under all schedulers a terminal state will eventually be reached a.s.
- if a cause candidate $\mathfrak{C}$ is given: 4 types of terminal states
  - ⋆ covered effect states: only accessible via $\mathfrak{C}$ (TP)

TP: true positive

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components
  i.e., under all schedulers a terminal state will eventually be reached a.s.
- if a cause candidate C is given: 4 types of terminal states
  - covered effect states: only accessible via C (TP)
  - uncovered effect states: not accessible from C (FN)

TP: true positive FN: false negative

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components
  i.e., under all schedulers a terminal state will eventually be reached a.s.
- if a cause candidate C is given: 4 types of terminal states
  - covered effect states: only accessible via C (TP)
  - uncovered effect states: not accessible from C (FN)
  - noneffect terminal states after C: only accessible via C (FP)

TP: true positive FN: false negative FP: false positive

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components
  i.e., under all schedulers a terminal state will eventually be reached a.s.
- if a cause candidate C is given: 4 types of terminal states
  - covered effect states: only accessible via C (TP)
  - uncovered effect states: not accessible from C (FN)
  - noneffect terminal states after C: only accessible via C (FP)
  - other noneffect terminal states: not accessible from C (TN)

TP: true positive FN: false negative FP: false positive TN: true negative

Model transformations

All algorithms rely on cause-effect preserving transformations to translate the original MDP into an equivalent one:

- with a single initial state and without end components

  i.e., under all schedulers a terminal state will eventually be reached a.s.

- if a cause candidate C is given: 4 types of terminal states

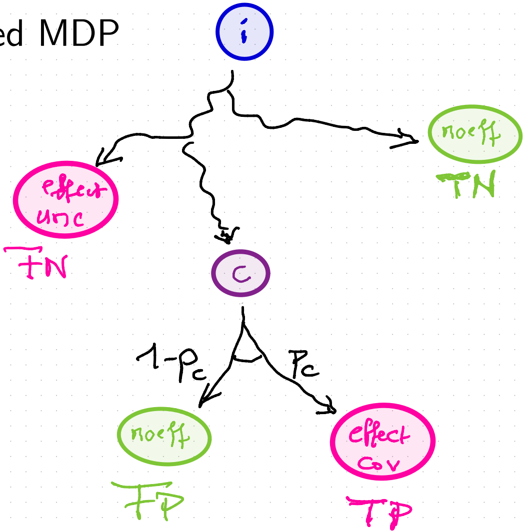
  - ⋆ covered effect states: only accessible via C (TP)
  - ⋆ uncovered effect states: not accessible from C (FN)
  - ⋆ noneffect terminal states after C: only accessible via C (FP)
  - ⋆ other noneffect terminal states: not accessible from C (TN)

  and each $c \in C$ has a single action with terminal successors

  (a covered effect state with prob. $p_c = \Pr_c^{\min}(\Diamond E)$ and a noneffect state with prob. $1 - p_c$)

Model transformation

Structure of the transformed MDP for fixed effect set *E* and cause candidate *C*:



Checking the SPR condition

Checking the SPR condition

Task: Given $\mathcal{M}$, $\boldsymbol{E}$, $\boldsymbol{C}$, check whether $\boldsymbol{C}$ is an SPR cause.

Checking the SPR condition

Task: Given $\mathcal{M}, E, C$, check whether C is an SPR cause.

Observation:

C is an SPR cause iff $\{c\}$ is an SPR cause for each state $c \in C$

Checking the SPR condition

Task: Given $\mathcal{M}$, E, C, check whether C is an SPR cause.

Observation:

C is an SPR cause iff $\{c\}$ is an SPR cause for each state $c \in C$

Existence of SPR or GPR causes:

there is an SPR cause

iff there is a singleton SPR cause

Checking the SPR condition

Task: Given $\mathcal{M}, E, C$, check whether C is an SPR cause.

Observation:

C is an SPR cause iff $\{c\}$ is an SPR cause for each state $c \in C$

Existence of SPR or GPR causes:

there is an SPR cause

iff there is a singleton SPR cause

iff there is a singleton GPR cause

iff there is a GPR cause

Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

Observation:

C is an SPR cause iff $\{c\}$ is an SPR cause for each state $c \in C$

Existence of SPR or GPR causes:

there is an SPR cause

iff there is a singleton SPR cause

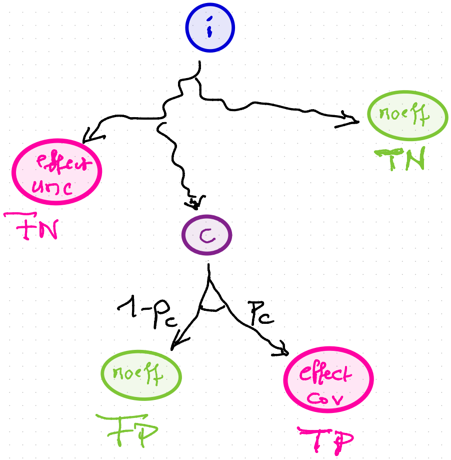
iff there is a singleton GPR cause

iff there is a GPR cause

Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

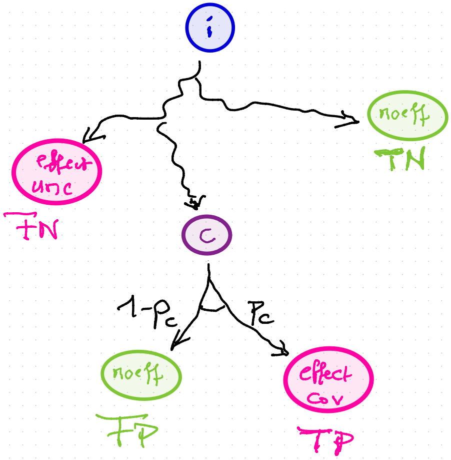
Let $\mathcal{N}$ be the transformed MDP where $p_c = \Pr^{\min}_{\mathcal{M},c}(\Diamond E)$.



Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP
where $p_c = \Pr^{\min}_{\mathcal{M},c}(\Diamond E)$.



$$p_c = \Pr^{\sigma}_{\mathcal{N},c}(\ \Diamond E \mid \Diamond c\ )$$

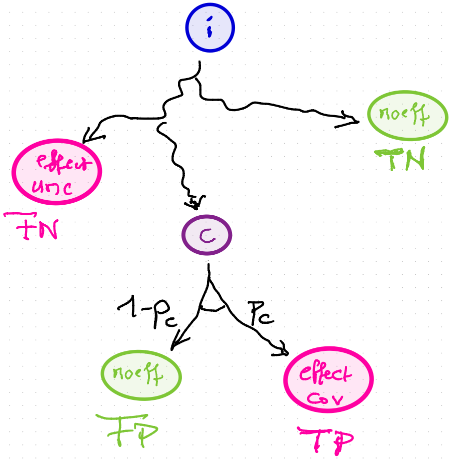
for each scheduler σ in $\mathcal{N}$
that reaches c

Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP
where $p_c = \Pr^{\min}_{\mathcal{M},c}(\lozenge E)$.

Let $q = \Pr^{\max}_{\mathcal{N}}(\lozenge E)$.



$p_c = \Pr^{\sigma}_{\mathcal{N},c}(\ \lozenge E \mid \lozenge c\ )$
for each scheduler σ in $\mathcal{N}$
that reaches c

Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

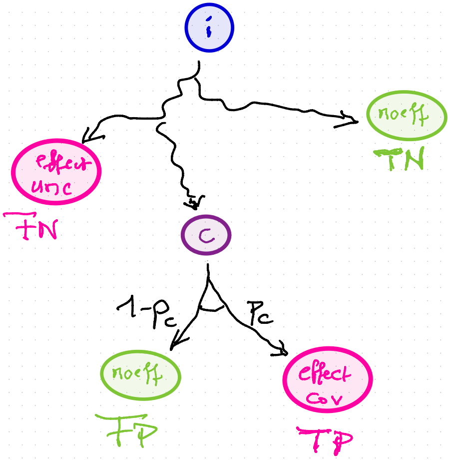
Let $\mathcal{N}$ be the transformed MDP
where $p_c = \Pr^{\min}_{\mathcal{M},c}(\Diamond E)$.

Let $q = \Pr^{\max}_{\mathcal{N}}(\Diamond E)$.

If $q < p_c$: SPR condition holds.

If $q > p_c$: SPR condition does not hold.

$p_c = \Pr^{\sigma}_{\mathcal{N},c}(\ \Diamond E \mid \Diamond c\ )$
for each scheduler σ in $\mathcal{N}$
that reaches c



Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP where $p_c = \Pr^{\min}_{\mathcal{M},c}(\lozenge E)$.

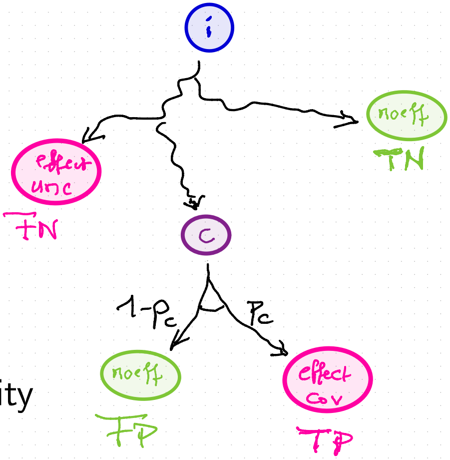
Let $q = \Pr^{\max}_{\mathcal{N}}(\lozenge E)$.

If $q < p_c$: SPR condition holds.

If $q > p_c$: SPR condition does not hold.

If $q = p_c$:

SPR condition holds iff $\mathcal{M}$ has no scheduler maximizing the effect probability that reaches c



Checking the SPR condition for singletons

Task: Given $\mathcal{M}$, E, c, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP
where $p_c = \Pr^{\min}_{\mathcal{M},c}(\Diamond E)$.

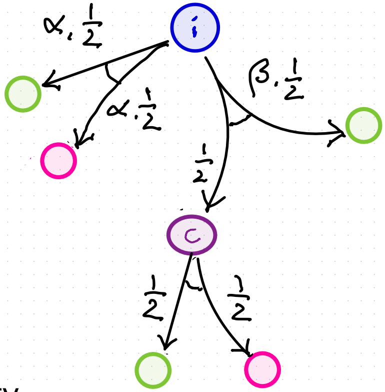
Let $q = \Pr^{\max}_{\mathcal{N}}(\Diamond E)$.

If $q < p_c$: SPR condition holds.

If $q > p_c$: SPR condition does not hold.

If $q = p_c$:

SPR condition holds iff $\mathcal{M}$ has no scheduler maximizing the effect probability that reaches c



Checking the GPR condition

Checking the GPR condition

After the model transformation:

$$C \text{ violates the GPR condition} \quad \text{iff} \quad \left\{ \begin{array}{l} \text{there is an MR-scheduler} \\ \text{refuting the GPR condition} \end{array} \right.$$

Checking the GPR condition

After the model transformation:

$$\mathcal{C} \text{ violates the GPR condition} \quad \text{iff} \quad \left\{ \begin{array}{l} \text{there is an MR-scheduler} \\ \text{refuting the GPR condition} \end{array} \right.$$

Main idea:

use a constraint system with variables

x_s for the expected frequencies of states $s \in S$, and

$x_{s,\alpha}$ for the expected frequencies of state-action pairs (s, α)

under such an MR-scheduler violating the GPR condition

Checking the GPR condition

- linear balance equations for the expected frequencies:

$$
\begin{aligned}
x_t &= \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,t) && \text{for each non-initial state } t \\
x_{s_0} &= \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,s_0) + 1 && \text{for the initial state } s_0
\end{aligned}
$$

Checking the GPR condition

- linear balance equations for the expected frequencies:

$$x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,t) \qquad \text{for each non-initial state } t$$

$$x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,s_0) + 1 \quad \text{for the initial state } s_0$$

- quadratic constraint for the violation of the GPR-condition:

$$x_C \cdot x_{FN} \geqslant (1 - x_C) \cdot \sum_{s \in C} x_s \cdot p_s$$

where $x_C = \sum_{s \in C} x_s$ (probability for reaching C), $p_s = \mathrm{Pr}_s^{\min}(\Diamond E)$ and

$x_{FN} = \sum_{s \in FN} x_s$ (prob. for false negatives, i.e., effect without cause)

Checking the GPR condition

- linear balance equations for the expected frequencies:

$$x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,t) \qquad \text{for each non-initial state } t$$

$$x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,s_0) + 1 \quad \text{for the initial state } s_0$$

- quadratic constraint for the violation of the GPR-condition:

$$x_C \cdot x_{FN} \geqslant (1 - x_C) \cdot \sum_{s \in C} x_s \cdot p_s$$

where $x_C = \sum_{s \in C} x_s$ (probability for reaching C), $p_s = \Pr^{\min}_s(\Diamond E)$ and

$x_{FN} = \sum_{s \in FN} x_s$ (prob. for false negatives, i.e., effect without cause)

$$\Pr(\Diamond E \mid \Diamond C) = \frac{\sum_{s \in C} x_s \cdot p_s}{x_C}$$

Checking the GPR condition

- linear balance equations for the expected frequencies:

$$x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,t) \qquad \text{for each non-initial state } t$$

$$x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,s_0) + 1 \quad \text{for the initial state } s_0$$

- quadratic constraint for the violation of the GPR-condition:

$$x_C \cdot x_{FN} \geqslant (1-x_C) \cdot \sum_{s \in C} x_s \cdot p_s$$

where $x_C = \sum_{s \in C} x_s$ (probability for reaching C), $p_s = \Pr^{\min}_s(\Diamond E)$ and

$x_{FN} = \sum_{s \in FN} x_s$ (prob. for false negatives, i.e., effect without cause)

$$\Pr(\Diamond E \mid \Diamond C) = \frac{\sum_{s \in C} x_s \cdot p_s}{x_C} \quad \text{and} \quad \Pr(\Diamond E \mid \neg\Diamond C) = \frac{x_{FN}}{1-x_C}$$

Checking the GPR condition

- linear balance equations for the expected frequencies:

$$x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,t) \qquad \text{for each non-initial state } t$$
$$x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s,\alpha,s_0) + 1 \quad \text{for the initial state } s_0$$

- quadratic constraint for the violation of the GPR-condition:

$$x_C \cdot x_{FN} \geqslant (1 - x_C) \cdot \sum_{s \in C} x_s \cdot p_s$$

where $x_C = \sum_{s \in C} x_s$ (probability for reaching C), $p_s = \Pr_s^{\min}(\Diamond E)$ and

$x_{FN} = \sum_{s \in FN} x_s$ (prob. for false negatives, i.e., effect without cause)

- linear non-negativity and positivity constraints:

$x_C > 0$ and $x_{s,\alpha} \geqslant 0$ for all state-action pairs

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}$, E, C, check whether

- C is an SPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: poly-time by statewise checking of the SPR condition

- C is a GPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: in PSPACE, using an encoding of the violation of the GPR condition in ETR (quadratic + linear constraints)

Finding good causes for given effects: Given $\mathcal{M}$, E, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Algorithmic problems

Checking cause-effect relationships: Given $\mathcal{M}, E, C$, check whether

- C is an SPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: poly-time by statewise checking of the SPR condition

- C is a GPR cause for E

  MC: poly-time using standard methods for (conditional) probabilities

  MDP: in PSPACE, using an encoding of the violation of the GPR condition in ETR (quadratic + linear constraints)

Finding good causes for given effects: Given $\mathcal{M}, E$, determine a PR cause C that is optimal w.r.t. to some coverage criterion.

Quality measures for causes

Quality measures for causes

- for fixed effect set E and GPR cause C
- take inspiration of quality measures used in statistical analysis for a good coverge of effect scenarios

Quality measures for causes

- for fixed effect set $\boldsymbol{E}$ and GPR cause $\boldsymbol{C}$
- take inspiration of quality measures used in statistical analysis for a good coverge of effect scenarios
- algorithmic problems:
  - ⋆ compute quality measure for fixed effect and GPR cause
  - ⋆ find optimal GPR cause for fixed effect set

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond C) \qquad \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}$$

ranges over all schedulers
with $\Pr^{\sigma}_{\mathcal{M}}(\Diamond C) > 0$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond E \mid \Diamond C\ )$$

$$\frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}$$

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond C \mid \Diamond E\ )$$

$$\frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}}$$

ranges over all schedulers
with $\Pr^{\sigma}_{\mathcal{M}}(\Diamond E) > 0$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond E \mid \Diamond C\ ) \qquad \frac{TP}{TP + FP}$$

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond C \mid \Diamond E\ ) \qquad \frac{TP}{TP + FN}$$

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C)\, \mathsf{U}\, E)} \qquad \frac{TP}{FN}$$

ranges over all schedulers
with $\Pr^{\sigma}_{\mathcal{M}}((\neg C)\, \mathsf{U}\, E) > 0$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond C) \qquad \frac{TP}{TP + FP}$$

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond C \mid \Diamond E) \qquad \frac{TP}{TP + FN}$$

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C) \, \mathrm{U} \, E)} \qquad \frac{TP}{FN}$$

f-score (harmonic mean of precision and recall)

$$fscore(C) = \inf_{\sigma} \frac{prec^{\sigma}(C) \cdot recall^{\sigma}(C)}{prec^{\sigma}(C) + recall^{\sigma}(C)}$$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond C)$$

already taken into account in the GPR condition;
precision says nothing about coverage

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond C \mid \Diamond E)$$

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C)\,\mathrm{U}\,E)}$$

$$\frac{\mathrm{TP}}{\mathrm{FN}}$$

f-score (harmonic mean of precision and recall)

$$fscore(C) = \inf_{\sigma} \frac{prec^{\sigma}(C) \cdot recall^{\sigma}(C)}{prec^{\sigma}(C) + recall^{\sigma}(C)}$$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond E \mid \Diamond C\ )$$

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\ \Diamond C \mid \Diamond E\ )$$

computing precision & recall:

via standard techniques for condition prob. in MDPs

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C)\,\mathsf{U}\,E)} \qquad \frac{\mathrm{TP}}{\mathrm{FN}}$$

f-score (harmonic mean of precision and recall)

$$fscore(C) = \inf_{\sigma} \frac{prec^{\sigma}(C) \cdot recall^{\sigma}(C)}{prec^{\sigma}(C) + recall^{\sigma}(C)}$$

Quality measures for causes

precision (accuracy for "(true or false) positives")

$$prec(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond E \mid \Diamond C)$$

recall (sensitivity):

$$recall(C) = \inf_{\sigma} \Pr^{\sigma}_{\mathcal{M}}(\Diamond C \mid \Diamond E)$$

computing precision & recall:

via standard techniques for condition prob. in MDPs

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C) \, \mathrm{U} \, E)}$$

f-score (harmonic mean of precision and recall)

$$fscore(C) = \inf_{\sigma} \frac{prec^{\sigma}(C) \cdot recall^{\sigma}(C)}{prec^{\sigma}(C) + recall^{\sigma}(C)}$$

computing covrat & f-score:

via reduction to SSPP (stoch. shortest path problem)

Coverage ratio and f-score

Coverage ratio and f-score

coverage ratio (fraction of covered and uncovered effects)

$$\mathit{covrat}(C) \;=\; \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C) \,\mathsf{U}\, E)} \;=\; \inf_{\sigma} \frac{\mathrm{TP}^{\sigma}}{\mathrm{FN}^{\sigma}}$$

f-score (harmonic mean of precision and recall)

$$\mathit{fscore}(C) = \inf_{\sigma} \frac{\mathit{prec}^{\sigma}(C) \cdot \mathit{recall}^{\sigma}(C)}{\mathit{prec}^{\sigma}(C) + \mathit{recall}^{\sigma}(C)}$$

TP true positive (covered effects)
FN false negative (uncovered effects)

Coverage ratio and f-score

coverage ratio (fraction of covered and uncovered effects)

$$\mathit{covrat}(C) \;=\; \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C) \,\mathsf{U}\, E)} \;=\; \inf_{\sigma} \frac{\mathsf{TP}^{\sigma}}{\mathsf{FN}^{\sigma}}$$

f-score (harmonic mean of precision and recall)

$$\mathit{fscore}(C) = \frac{2}{X+2} \quad \text{where} \quad X = \sup_{\sigma} \frac{\mathsf{FP}^{\sigma} + \mathsf{FN}^{\sigma}}{\mathsf{TP}^{\sigma}}$$

TP true positive (covered effects)
FN false negative (uncovered effects)

TN true negative (noeffect without C)
FP false positive (noffect after C)

Coverage ratio and f-score

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) \;=\; \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C)\,\mathrm{U}\,E)} \;=\; \inf_{\sigma} \frac{\mathrm{TP}^{\sigma}}{\mathrm{FN}^{\sigma}}$$

f-score (harmonic mean of precision and recall)

$$fscore(C) = \frac{2}{X+2} \quad \text{where} \quad X = \sup_{\sigma} \frac{\mathrm{FP}^{\sigma} + \mathrm{FN}^{\sigma}}{\mathrm{TP}^{\sigma}}$$

After model transformation for fixed effect and GPR cause:

- TP, FP, FN, TN are terminal states

Coverage ratio and f-score

coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond C \wedge \Diamond E)}{\Pr^{\sigma}_{\mathcal{M}}((\neg C) \, \mathsf{U} \, E)} = \inf_{\sigma} \frac{\mathsf{TP}^{\sigma}}{\mathsf{FN}^{\sigma}}$$

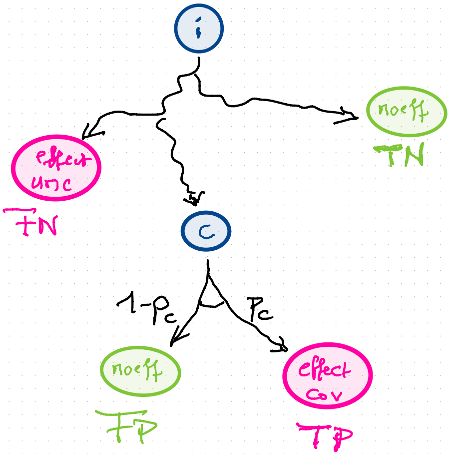
f-score (harmonic mean of precision and recall)

$$fscore(C) = \frac{2}{X+2} \quad \text{where} \quad X = \sup_{\sigma} \frac{\mathsf{FP}^{\sigma} + \mathsf{FN}^{\sigma}}{\mathsf{TP}^{\sigma}}$$

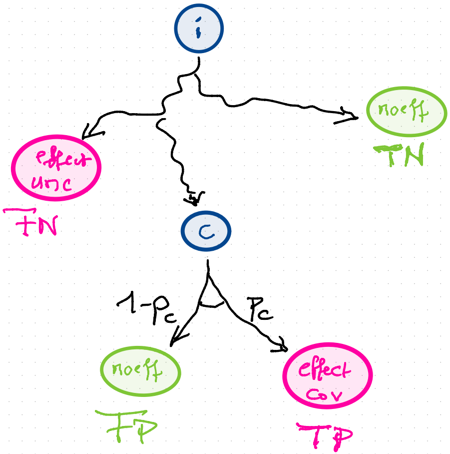
After model transformation for fixed effect and GPR cause:

- TP, FP, FN, TN are terminal states
- recall and f-score can be derived from inf resp. sup of $\frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$
  quotient of probabilities for reaching disjoint sets of terminal states

After model transformation ...



After model transformation ...

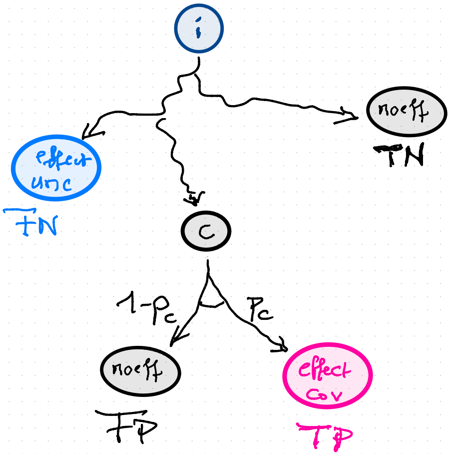


coverage ratio

fraction of covered and uncovered effects

$$covrat(C) = \inf_{\sigma} \frac{\mathrm{TP}^{\sigma}}{\mathrm{FN}^{\sigma}}$$

After model transformation ...

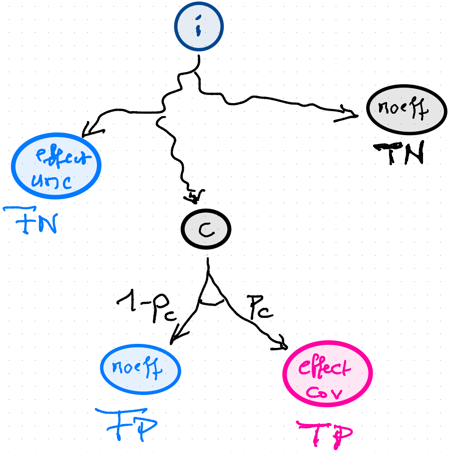


coverage ratio

fraction of covered and uncovered effects

$$covrat(C) = \inf_{\sigma} \frac{TP^{\sigma}}{FN^{\sigma}}$$

After model transformation ...



coverage ratio

fraction of covered and uncovered effects

$$\textit{covrat}(C) = \inf_{\sigma} \frac{\mathrm{TP}^{\sigma}}{\mathrm{FN}^{\sigma}}$$

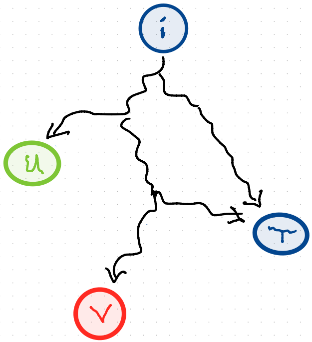
f-score

harmonic mean of precision & recall

$$\textit{fscore}(C) = \frac{2}{X+2}$$

where $X = \sup_{\sigma} \dfrac{\mathbf{FP}^{\sigma} + \mathbf{FN}^{\sigma}}{\mathbf{TP}^{\sigma}}$

Covratio and f-score via SSPP



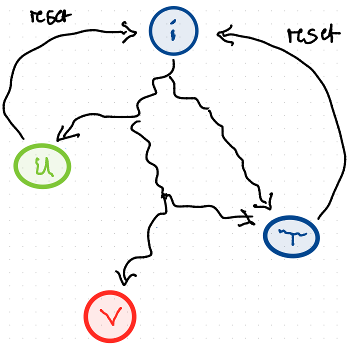
Given MDP $\mathcal{M}$

- without end components
- U, V disjoint sets of terminal states

Goal: compute $\inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$

(for sup analogous)

Covratio and f-score via SSPP



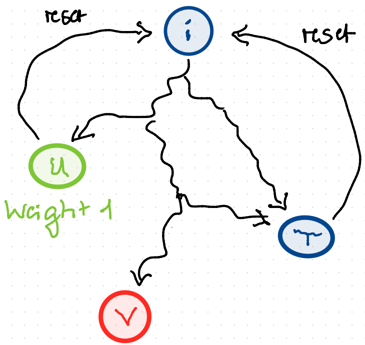
Given MDP $\mathcal{M}$

- without end components
- U, V disjoint sets of terminal states

Goal: compute $\inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\lozenge U)}{\Pr^{\sigma}_{\mathcal{M}}(\lozenge V)}$

(for sup analogous)

Covratio and f-score via SSPP



Given MDP $\mathcal{M}$

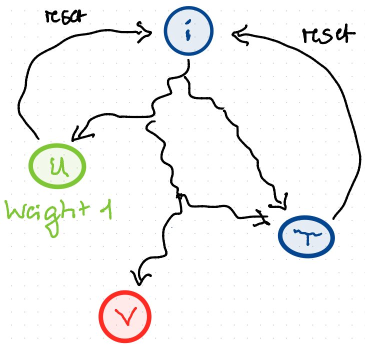
- without end components
- U, V disjoint sets of terminal states

Goal: compute $\inf\limits_{\sigma} \dfrac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP

weight 1 for U, weight 0 for all other states

Covratio and f-score via SSPP



stochastic process

initially: $w = 0$

Given MDP $\mathcal{M}$

- without end components
- U, V disjoint sets of terminal states

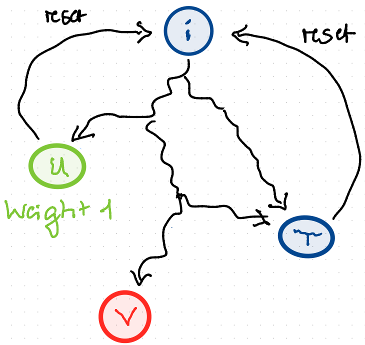
Goal: compute $\inf\limits_{\sigma} \dfrac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP

weight 1 for U, weight 0 for all other states

1. generate sample run until reaching a terminal state s
2. If $s \in V$ then return w and halt.
   If $s \in U$ then $w := w+1$ and go to 1.
   If $s \in T$ (other terminal state) then go to 1.

Covratio and f-score via SSPP



Given MDP $\mathcal{M}$

- without end components
- U, V disjoint sets of terminal states

Goal: compute $\inf\limits_{\sigma} \dfrac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP

weight 1 for U, weight 0 for all other states

stochastic process

initially: $w = 0$

expected outcome:

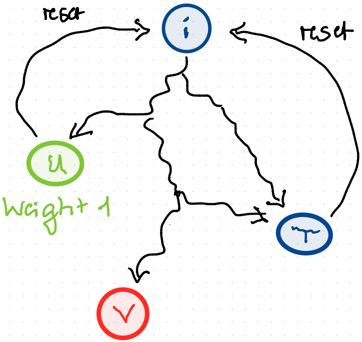
$$\frac{\Pr(\Diamond U)}{\Pr(\Diamond V)}$$

1. generate sample run until reaching a terminal state s
2. If $s \in V$ then return w and halt.

   If $s \in U$ then $w := w+1$ and go to 1.

   If $s \in T$ (other terminal state) then go to 1.

Covratio and f-score via SSPP



Given MDP $\mathcal{M}$

- without end components
- U, V disjoint sets of terminal states

Goal: compute $\inf\limits_{\sigma} \dfrac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP

weight 1 for U, weight 0 for all other states

$$\inf_{\sigma} \frac{\Pr^{\sigma}_{\mathcal{M}}(\Diamond U)}{\Pr^{\sigma}_{\mathcal{M}}(\Diamond V)} \;=\; \inf_{\sigma} \mathbb{E}^{\sigma}_{\mathcal{N}}(\text{“accumulated weight until reaching } V\text{”})$$

stochastic shortest path in $\mathcal{N}$

Quality measures for causes

- Three measures for the *"degree of coverage"*:

  recall, coverage ratio, and f-score

- computable in poly-time for fixed effect E and GPR cause C:
  - recall: via standard techniques for conditional probabilities in MDPs
  - coverage ratio and f-score: via polynomial reduction to SSPP

Quality measures for causes

- Three measures for the *"degree of coverage"*:

  recall, coverage ratio, and f-score

- computable in poly-time for fixed effect E and GPR cause C:
  - ⋆ recall: via standard techniques for conditional probabilities in MDPs
  - ⋆ coverage ratio and f-score: via polynomial reduction to SSPP

- optimalization problem:
  given effect set E, find an SPR or a GPR cause C with
  - ⋆ maximal recall
  - ⋆ maximal coverage ratio
  - ⋆ maximal f-score

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

Optimal SPR causes:

- recall-optimal = covratio-optimal: computable in poly-time

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

Optimal SPR causes:

- recall-optimal = covratio-optimal: computable in poly-time

"canonical SPR cause": $\mathcal{C}$ = union of all singleton SPR causes

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- ⋆ in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

Optimal SPR causes:

- ⋆ recall-optimal = covratio-optimal: computable in poly-time

"canonical SPR cause": $\mathcal{C}$ = union of all singleton SPR causes

- recall-optimal: obvious as any SPR is a subset of $\mathcal{C}$

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

Optimal SPR causes:

- recall-optimal = covratio-optimal: computable in poly-time

"canonical SPR cause": $\mathcal{C}$ = union of all singleton SPR causes

- recall-optimal: obvious as any SPR is a subset of $\mathcal{C}$
- covratio-opt = recall-opt: $\frac{\text{TP}}{\text{FN}} < \frac{\text{TP'}}{\text{FN'}}$ iff $\frac{\text{TP}}{\text{FN+TP}} < \frac{\text{TP'}}{\text{FN'+TP'}}$

Finding optimal causes

Optimal GPR causes (recall, coverage ratio and f-score):

- ⋆ in polynomial space

  by considering all cause candidates, checking the GPR condition (poly-space) and computing their recall, coverage ratio or f-score (poly-time)

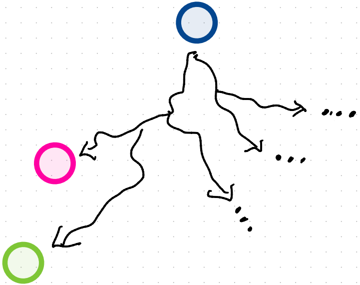
Optimal SPR causes:

- ⋆ recall-optimal = covratio-optimal: computable in poly-time
- ⋆ f-score optimal causes:

  MC: in poly-time via reduction to SSPP in MDPs

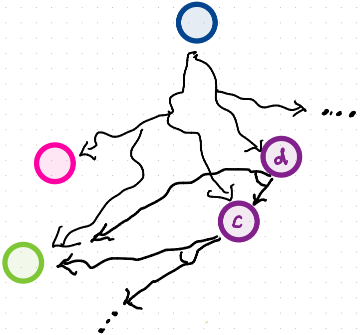
  MDP: in exp-time via reduction to SSP-games

F-score optimal SPR cause in MC



MC $\mathcal{M}$

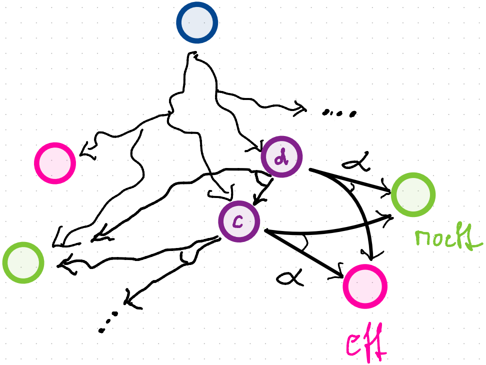
F-score optimal SPR cause in MC



MC $\mathcal{M}$

$\mathcal{C} = \{c, d, \ldots\}$
set of states c with
$p_c = \Pr_c(\Diamond E) > \Pr_{\mathcal{M}}(\Diamond E)$

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

$\mathcal{C} = \{c, d, \dots\}$
set of states c with
$p_c = \Pr_c(\lozenge E) > \Pr_{\mathcal{M}}(\lozenge E)$

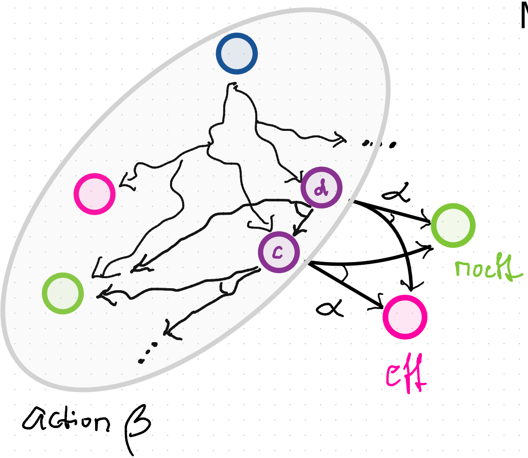
nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"

move with prob. p_c to new effect state *eff*

with prob. $1 - p_c$ to a terminal non-effect state

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

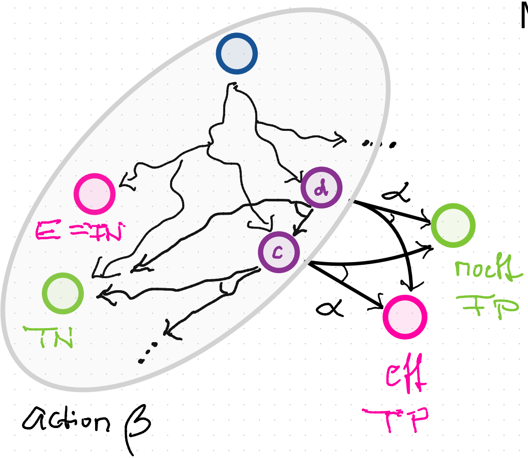
$\mathcal{C} = \{c, d, \ldots\}$
set of states c with
$p_c = \Pr_c(\Diamond E) > \Pr_{\mathcal{M}}(\Diamond E)$

nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"
  move with prob. p_c to new effect state *eff*
  with prob. $1{-}p_c$ to a terminal non-effect state

action β: "c not selected for SPR cause"

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

$\mathcal{C} = \{c, d, \ldots\}$
set of states c with
$p_c = \Pr_c(\lozenge E) > \Pr_{\mathcal{M}}(\lozenge E)$

nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"

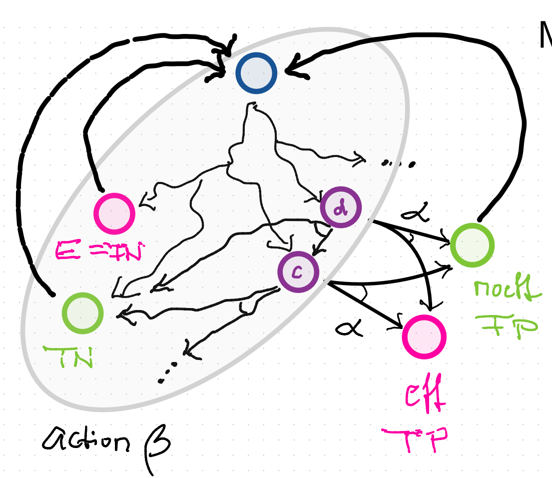
move with prob. p_c to new effect state *eff*

with prob. $1 - p_c$ to a terminal non-effect state

action β: "c not selected for SPR cause"

$$fscore(C) = \frac{2}{X_C + 2} \quad \text{where} \quad X_C = \frac{FN_C + FP_C}{TP_C}$$

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

$\mathcal{C} = \{c, d, ...\}$
set of states c with
$p_c = \Pr_c(\Diamond E) > \Pr_{\mathcal{M}}(\Diamond E)$

nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"

move with prob. p_c to new effect state *eff*

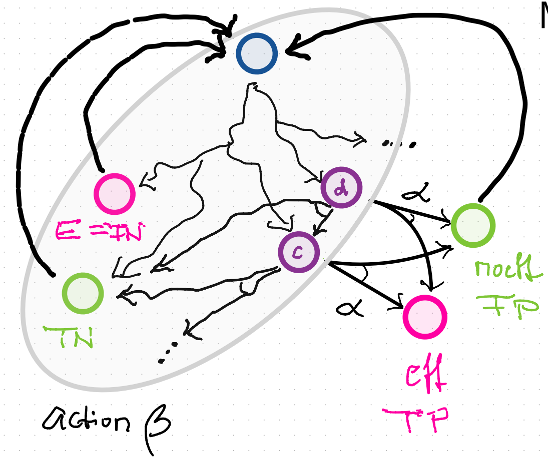
with prob. $1-p_c$ to a terminal non-effect state

action β: "c not selected for SPR cause"

reset transitions from TP, FN, FP

$fscore(C) = \frac{2}{X_C+2}$ where $X_C = \frac{FN_C+FP_C}{TP_C}$

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

$\mathcal{C} = \{c, d, \ldots\}$
set of states c with
$p_c = \Pr_c(\Diamond E) > \Pr_{\mathcal{M}}(\Diamond E)$

nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"

move with prob. p_c to new effect state *eff*

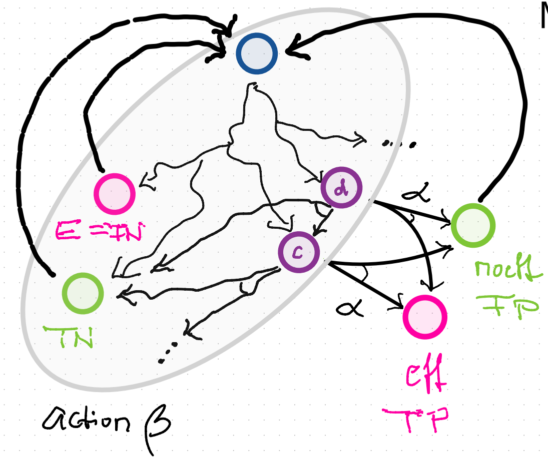
with prob. $1-p_c$ to a terminal non-effect state

action β: "c not selected for SPR cause"

reset transitions from TP, FN, FP
weight 1 for FN and FP
weight 0 for all other states

$\mathit{fscore}(C) = \frac{2}{X_C+2}$ where $X_C = \frac{\mathrm{FN}_C+\mathrm{FP}_C}{\mathrm{TP}_C}$

F-score optimal SPR cause in MC



MDP $\mathcal{N}$

$\mathcal{C} = \{c, d, \ldots\}$
set of states c with
$p_c = \Pr_c(\Diamond E) > \Pr_{\mathcal{M}}(\Diamond E)$

nondeterministic choice in $\mathcal{C}$-states

action α: "c selected for SPR cause"

move with prob. p_c to new effect state *eff*

with prob. $1-p_c$ to a terminal non-effect state

action β: "c not selected for SPR cause"

reset transitions from TP, FN, FP
weight 1 for FN and FP
weight 0 for all other states

$$\max_{C} \mathit{fscore}(C) = \frac{2}{X+2} \quad \text{where } X = \mathbb{E}^{\min}_{\mathcal{N}}(\text{weight})$$

Summary: algorithmic problems for PR causes

Summary: algorithmic problems for PR causes

Results on strict and global probability-raising causality in Markov chains and MDPs (with fixed effect set E):

For fixed set C:

	checking PR condition	computing quality measures (recall, coverage ratio, f-score)
SPR	$\in$ P	poly-time
GPR	MDP: $\in$ PSPACE MC: $\in$ P	poly-time

Summary: algorithmic problems for PR causes

Results on strict and global probability-raising causality in Markov chains and MDPs (with fixed effect set E):

Finding optimal causes and related threshold problems:

	covratio-optimal = recall-optimal	f-score-optimal	threshold problem
SPR	poly-time	MDP: poly-space MC: poly-time	f-score threshold problem MDP: $\in$ NP $\cap$ coNP MC: $\in$ P
GPR	poly-space		MDP: $\in$ PSPACE MC: NP-complete

Conclusions

Conclusions

part 1: notions of causality and responsibility in TS

- forward causality
  - ⋆ necessary and sufficient causes (formalization in CTL*)
  - ⋆ counterfactual: mutation- or game-based definition
    open: is there a logical characterization? (using some hyperlogic?)
- backward causality
  - ⋆ game-based definition of strategic and causal responsibility

Conclusions

part 1: notions of causality and responsibility in TS

- forward causality
  - ⋆ necessary and sufficient causes (formalization in CTL*)
  - ⋆ counterfactual: mutation- or game-based definition
    open: is there a logical characterization? (using some hyperlogic?)
- backward causality
  - ⋆ game-based definition of strategic and causal responsibility
- measures for the importance of states on temporal properties
  - ⋆ degree of responsibility for the satisfaction of properties:
    mutation- or game-based definition via size of smallest switching pairs
  - ⋆ Shapley values to measure the importance of states on the truth of path formulas
    - quantitative version of forward responsibility
    - analogous for strategic backward responsibility, but unclear for causal backward resp.
    - more difficult for branching-time logics [Mascle et al, LICS'21]

Conclusions

part 1: notions of causality and responsibility in TS

- forward causality
  - ⋆ necessary and sufficient causes (formalization in CTL*)
  - ⋆ counterfactual: mutation- or game-based definition
    open: is there a logical characterization? (using some hyperlogic?)
- backward causality
  - ⋆ game-based definition of strategic and causal responsibility
- measures for the importance of states on temporal properties
  - ⋆ degree of responsibility for the satisfaction of properties:
    mutation- or game-based definition via size of smallest switching pairs
  - ⋆ Shapley values to measure the importance of states on the truth of path formulas
- Aumann-Shapley values for models with continuous parameters

  e.g., to measure the impact of probability parameters in parametric Markov chains on reachability probabilities or expected costs [B., Funke, Majumdar, AAAI'21]

Conclusions

part 1: notions of causality and responsibility in TS

- forward causality
  - ⋆ necessary and sufficient causes (formalization in CTL*)
  - ⋆ counterfactual: mutation- or game-based definition
    open: is there a logical characterization? (using some hyperlogic?)
- backward causality
  - ⋆ game-based definition of strategic and causal responsibility
- measures for the importance of states on temporal properties

⋮

part 2: probabilistic causality in Markovian models

- MDP-formalization of the PR condition $\mathbf{Pr}(\mathit{effect}|\mathit{cause}) > \mathbf{Pr}(\mathit{effect}|\neg\,\mathit{cause})$
- many open questions: path events for causes and effects, other quality measures, backward causality, actionability, ...

Thank You