From verification to causality-based explications

Christel Baier
TU Dresden

Joint work with:

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

Classical verification task:
   given: a system model $\mathcal{M}$ and a specification $\phi$
   question: does $\mathcal{M}$ satisfy $\phi$?
From verification to explications

Classical verification task:

- given: a system model $\mathcal{M}$ and a specification $\phi$
- question: does $\mathcal{M}$ satisfy $\phi$?
- answer: yes or no
From verification to explications

Classical verification task:

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

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

answer: yes or no

counterexample

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

mathematical proof or certificate
counterexample
From verification to explications

Classical verification task:

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

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

answer: yes or no

mathematical proof or certificate

Explication task (in the verification context):

... should provide deeper insights why the specification holds or not
Classical verification task:

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

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

answer: yes or no

- mathematical proof
- or certificate
- counterexample

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 $M$ and a specification $\phi$

question: does $M$ satisfy $\phi$?

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 requirement violation and to which degree?
- if a bad behavior occurs, what has caused the violation of the specification?
Causality

long-standing discussion in philosophy

David Hume
(philosopher, 1711-1776)

David K. Lewis
(philosopher, 1941-2001)

and many more ...
Causality

long-standing discussion in philosophy, but also AI

Joseph Halpern
Gödel Prize 1997
Dijkstra Prize 2009

Judea Pearl
Turing Award
Winner 2011

© CC BY-SA 2.0 fr
Joe Halpern at EPFL in June 2008
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”

- 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 occurred

  *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...
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 occurred
  - *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 occurred
  - 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

- 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
  - Study mutations of system models or specifications [Chockler et al'01, Beer et al'01, Kupferman/Vardi'03]
- 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

which program statements affect the values of variables at a certain program location?
Causality in the verification context

- program slicing
- causality-based explanations of counterexamples
  - counterfactual reasoning with distance metrics
  - identification of “critical state-variable pairs” in cex
  - event order logic for causal dependencies in cex

[Weiser’79]

*[Groce et al’06]*

*[Beer et al’09]*

*[Leitner-Fischer/Leue’13]*

coverage and vacuity

study mutations of system models or specifications

causality and responsibility in operational models

cause-effect relations

quantitative measures for the relevance of states

proof rules for stepwise cause-effect reasoning

[Chockler et al’01, Beer et al’01, Kupferman/Vardi’03]

*[Cho./Hal./Kup.’08, B./Funke/Maj.’21, B./Funke/Pir./Zie.’22]*

*[Chockler/Halpern/Kupf.’08, B./Funke/Maj.’21, Mascle et al’21]*

causality-based verification

*[Kupriyanov/Finkbeiner’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./Funke/Maj.’21, B./Funke/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./Funke./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./Funke/Maj.’21, Mascle et al’21]
  - 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

Given a TS with state space $SSS$ and a set $E \subseteq SE \subseteq S$ 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:
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”
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. $C \cap E = \emptyset$. $C$ is called a

- necessary cause for $E$ if $\mathcal{M} \models \forall \neg((\neg C) \cup E)$
  
  “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. $C \cap E = \emptyset$. $C$ is called a

- necessary cause for $E$ if $\mathcal{M} \models \forall \neg((\neg C) \cup E)$
  
  “if the effect occurs then the cause must have happened before”

- sufficient cause for $E$ if $\mathcal{M} \models \forall \square(C \rightarrow \lozenge\lozenge E)$
  
  “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. $C \cap E = \emptyset$. $C$ is called a

- necessary cause for $E$ if $\mathcal{M} \models \forall \neg ((\neg C) \cup E)$
  “if the effect occurs then the cause must have happened before”

- sufficient cause for $E$ if $\mathcal{M} \models \forall \Box (C \rightarrow \Diamond \Diamond E)$
  “if the cause happens then the effect will occur somewhen later”

Monotonicity:

- $C$ is necessary and $C \subseteq D \implies D$ is necessary
- $C$ is sufficient and $C \supseteq D \implies D$ 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. $C \cap E = \emptyset$. $C$ is called a

- necessary cause for $E$ if $\mathcal{M} \models \forall \neg ((\neg C) \cup E)$
  "if the effect occurs then the cause must have happened before"

- sufficient cause for $E$ if $\mathcal{M} \models \forall \Box (C \rightarrow \Diamond \Diamond E)$
  "if the cause happens then the effect will occur somewhen later"

Transitivity (up to disjointness):

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

$C$ sufficient for $D$ & $D$ sufficient for $E$ $\implies$ $C$ 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. $C \cap E = \emptyset$. $C$ is called a

- necessary cause for $E$ if $\mathcal{M} \models \forall \neg ((\neg C) \cup E)$
  
  "if the effect occurs then the cause must have happened before"

- sufficient cause for $E$ if $\mathcal{M} \models \forall \Box (C \rightarrow \Diamond \Diamond E)$
  
  "if the cause happens then the effect will occur somewhen later"

If all $E$-states are terminal then:

- $C$ is necessary iff $\mathcal{M} \models \forall (\Diamond E \rightarrow \Diamond C)$
- $C$ is sufficient iff $\mathcal{M} \models \forall (\Diamond C \rightarrow \Diamond E)$
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. \( C \cap E = \emptyset \). Then:

\[
\begin{align*}
C \text{ is a necessary cause for } E & \iff \mathcal{M} \models \forall (\Diamond E \rightarrow \Diamond C) \\
C \text{ is a sufficient cause for } E & \iff \mathcal{M} \models \forall (\Diamond C \rightarrow \Diamond E)
\end{align*}
\]

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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$. 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
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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$. 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 $\mathcal{M} \models \forall (\Diamond C \rightarrow \Diamond E)$

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

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

$\lfloor C \rfloor$ results from $C$ by removing all states $s$ where each path $\pi$ 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. $C \cap E = \emptyset$. 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 $\mathcal{M} \models \forall (\Diamond C \rightarrow \Diamond E)$

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

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

... 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. $C \cap E = \emptyset$. 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. \( C \cap E = \emptyset \). Then:

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

Properties of sufficient causes:

- \( C_{E} \overset{\text{def}}{=} \{ s \in S : s \models \forall \Diamond \forall \Diamond E \} \) 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. $C \cap E = \emptyset$. Then:

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

Properties of sufficient causes:

- $C_E \overset{\text{def}}{=} \{ s \in S : s \models \forall \lozenge \forall \lozenge E \}$ is a sufficient cause ...
  and contains all other sufficient causes

- If $\mathcal{M} \not\models \exists \lozenge \forall \lozenge \forall \lozenge E$ then $\emptyset$ is the only sufficient cause.

there is no reachable state $s$
  s.t. $s \models \forall \lozenge \forall \lozenge E$
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. $C \cap E = \emptyset$. Then:

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

Properties of sufficient causes:

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

  ... and contains all other sufficient causes

- If $\mathcal{M} \nvDash \exists \Diamond \forall \Box \forall \Diamond E$ then $\emptyset$ is the only sufficient cause.

- Canonical sufficient cause: $\lfloor C_E \rfloor$

  pruning operator: $\lfloor C \rfloor = \{ s \in C : \mathcal{M} \models \exists (\neg C) \cup s \}$
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. $C \cap E = \emptyset$. Then:

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

Properties of sufficient causes:

- $C_E \overset{\text{def}}{=} \{s \in S : s \models \forall \lozenge \forall \lozenge E\}$ is a sufficient cause ... and contains all other sufficient causes

- If $\mathcal{M} \not\models \exists \lozenge \forall \lozenge \forall \lozenge E$ then $\emptyset$ is the only sufficient cause.

- Canonical sufficient cause: $[C_E]$
  
  ... 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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$. Then:

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

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 $E$. 


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. $C \cap E = \emptyset$. 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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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

**degree of sufficiency** ("precision")

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

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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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

The degree of sufficiency ("precision")

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

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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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

Degree of sufficiency ("precision")

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

Degree of necessity ("recall")

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

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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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

degree of sufficiency ("precision")

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

degree of necessity ("recall")

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

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. $C \cap E = \emptyset$ and $C, E \neq \emptyset$.

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

Degree of sufficiency ("precision")

$$\Pr_{\mathcal{M}}(\Diamond E | \Diamond C) = \frac{\Pr_{\mathcal{M}}(\Diamond C \land \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond C)}$$

Degree of necessity ("recall")

$$\Pr_{\mathcal{M}}(\Diamond C | \Diamond E) = \frac{\Pr_{\mathcal{M}}(\Diamond C \land \Diamond E)}{\Pr_{\mathcal{M}}(\Diamond E)}$$

$C$ and $\lceil C \rceil$ have the same degree of sufficiency 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:

$$[C_E] \text{ where } C_E \overset{\text{def}}{=} \{ s \in S : s \models \forall \Box \forall \Diamond E \}$$
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:

$$[C_E] \text{ where } C_E \overset{\text{def}}{=} \{ s \in S : s \models \forall \Box \forall \Diamond E \}$$

Necessary causes with maximal degree of sufficiency:

$$[\text{Pre}(E)]$$
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:

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

Necessary causes with maximal degree of sufficiency:

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

$$[C_E] \text{ where } C_E \overset{\text{def}}{=} \{ s \in S : s \models \forall \bigcirc \forall \bigtriangledown E \}$$

Necessary causes with maximal degree of sufficiency:

$$[\text{Pre}(E)] \text{ and } [C] \text{ where } C = \{ s \in S : \Pr_s(\bigtriangledown \text{Pre}(E)) = 1 \}$$

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
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

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 non-actual causality and nonspecific actual causality. These are sometimes called causal and specific causality. Type causality is a general statement, such as "smoking causes lung cancer," whereas actual causality focuses on particular events: "the accident (not the year - 62)"

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, 2003] to deal with some problems pointed out by Hopkins [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 (c) gives reasonable answers that agree with those of the original examples of causality and (d) has lower complexity than either the original or updated definitions.

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, shatters 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 Billy’s throw a cause of the bottle shattering.

More formally, we can use binary variables to model the world: \( S = 1 \) if Suzy throws, \( S = 0 \) if she doesn’t; \( B = 1 \) if Billy shatters the bottle; \( B = 0 \) if he doesn’t. Our definitions allow us to consider a situation where the HP definition doesn’t work as to determine causality.
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 occurred

  *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

**Structural equation model:**

\[ S = (\text{Exo}, \text{Endo}, f) \]

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

**Interventions:**

Given \( Y \subseteq \text{Endo} \) and \( \beta \in \text{Val}(Y) \), let \( S[Y \leftarrow \beta] \) when the \( Y \)-variables are treated as constants given by the values in \( \beta \) for counterfactual reasoning: “enforce values of endogenous variables (ignoring their equations)”
Structural equation model: \( S = (Exo, Endo, f) \) where

- **Exo**: set of exogenous variables (specify the context)
- **Endo**: totally ordered set of endogenous variables, say \( x_1, \ldots, x_n \)

\[
\begin{align*}
x_1 & \text{ only depends on the context} \\
x_2 & \text{ only depends on the context and } x_1 \\
x_3 & \text{ only depends on the context and } x_1, x_2 \\
\vdots & \ddots
\end{align*}
\]
HP structural equation model

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

$\mathbf{Exo}$: set of exogenous variables (specify the context)

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

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

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

$x_1$ only depends on the context

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

\vdots

$Val(\mathcal{V}) = \text{set of valuations for the variables in } \mathcal{V} \subseteq \mathbf{Exo} \cup \mathbf{Endo}$
Structural equation model: $\mathbf{S} = (\mathbf{Exo}, \mathbf{Endo}, \mathbf{f})$ where

$\mathbf{Exo}$: set of exogenous variables (specify the context)

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

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

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

\[
\begin{align*}
\alpha_1 &= S_1(c) \overset{\text{def}}{=} f_1(c) & (\text{value for } x_1) \\
\alpha_2 &= S_2(c) \overset{\text{def}}{=} f_2(c, \alpha_1) & (\text{value for } x_2) \\
&\vdots & \vdots \\
\alpha_n &= S_n(c) \overset{\text{def}}{=} f_n(c, \alpha_1, \ldots, \alpha_{n-1}) & (\text{value for } x_n)
\end{align*}
\]
HP structural equation model

Structural equation model: $S = (Exo, Endo, f)$ 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 $f_i : Val(Exo, x_1, \ldots, x_{i-1}) \rightarrow Val(x_i)$

Interventions:

for counterfactual reasoning:

“enforce values of endogenous variables (ignoring their equations)”
HP structural equation model

Structural equation model: \[ S = (\text{Exo}, \text{Endo}, f) \]

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

Interventions: given \( Y \subseteq \text{Endo} \) and \( \beta \in \text{Val}(Y) \), let
  \[ S[Y \leftarrow \beta] = \begin{cases} S & \text{when the } Y\text{-variables are treated as constants given by the values in } \beta \\ S & \text{otherwise} \end{cases} \]

for counterfactual reasoning:
  “enforce values of endogenous variables (ignoring their equations)”
Let $S = (\text{Exo}, \text{Endo}, f)$ be a structural equation model and

\[ \phi \in \text{Val} (\text{Exo}) \]

a context s.t. $(S, c) \models \phi$. Then $X = \alpha$ in context $c$ iff

\[ \neg \phi \]

$\subseteq$ Endo and $\alpha = S_X(c)$

\[ \neg \phi \]

\[ \neg \phi \]

\[ \neg \phi \]

... counterfactual condition ...

... minimality condition ....

tuple of values for $XXX$ in $SSS$ for context $ccc$ obtained by the equations

\[ x_i = f_i(c, x_1, \ldots, x_{i-1}) \]
HP causality

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

- $\varphi$ be a Boolean condition for the values of variables (exo or endo)
- $c \in \text{Val(Exo)}$ a context s.t. $(S, c) \models \varphi$
Let $S = (\text{Exo}, \text{Endo}, f)$ be a structural equation model and

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

tuple of values for $X$ in $S$ for context $c$
obtained by the equations

$x_i = f_i(c, x_1, \ldots, x_{i-1})$
Let $S = (\text{Exo}, \text{Endo}, f)$ be a structural equation model and

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

Then $X = \alpha$ is called a cause for $\varphi$ in context $c$ iff

1. $[\text{AC1}]$ ... counterfactual condition ...
2. $[\text{AC2}]$ ... minimality condition ....
Let $S = (\text{Exo}, \text{Endo}, f)$ be a structural equation model and

- $\phi$ be a Boolean condition for the values of variables (exo or endo)
- $c \in \text{Val}(\text{Exo})$ a context s.t. $(S, c) \models \phi$
- $X \subseteq \text{Endo}$ and $\alpha = S_X(c)$

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

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

$$(S[X \leftarrow \beta], c) \models \neg \phi$$

[AC2] ... minimality condition ....
Let $S = (\text{Exo}, \text{Endo}, f)$ be a structural equation model and

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

Then $X = \alpha$ is an actual cause for $\varphi$ in context $c$ iff

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

$(S[X \leftarrow \beta, Y \leftarrow S_Y(c)], c) \models \neg \varphi$

[AC2] ... minimality condition ....
HP causality

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

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

Then $X = \alpha$ is an actual cause for $\varphi$ in context $c$ iff

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

$$(S[X \leftarrow \beta, Y \leftarrow S_Y(c)], c) \models \neg \varphi$$

[AC2] $X$ is minimal w.r.t. condition [AC1]
HP causality and degree of responsibility

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

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

Then, the degree of responsibility of $x = \alpha$ for $\varphi$ is . . .

[Chockler/Halpern/Kupferman, ACM ToCL 2008]
HP causality and degree of responsibility

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

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

Then, the degree of responsibility of $x = \alpha$ for $\varphi$ 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]
Let $S = (Exo, Endo, f)$ be a structural equation model and

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

Then, the degree of responsibility of $x = \alpha$ for $\varphi$ is $\frac{1}{m}$ where $m$ is the minimal number of value-changes for endo variables required to make $\varphi$ counterfactually depend on $x$.

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

$$(S[X \leftarrow \beta, Y \leftarrow S_Y(c)], c) \models \neg \varphi$$
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
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 already proven to be effective in practice, detecting design errors that escape early verification efforts in industrial settings. In this paper, we relate a formal definition of causality given by Halpern and Pearl [2005] to coverage. We show that it gives significant insight into unresolved issues regarding the definition of coverage and leads to potentially useful extensions of coverage. In particular, we introduce the notion of responsibility, which assigns to components of a system a quantitative measure of their relevance to the satisfaction of the specification.
Counterfactual causality: backward vs forward

- **Backward counterfactual causality**: Given an effect scenario:
  
  "If the cause would not have happened, then the effect would not have occurred."

- **Forward counterfactual causality**: 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.
Counterfactuality: backward vs forward

backward counterfactual causality

given an effect scenario:

“if the cause would not have happened, then the effect would not have occurred”

intervention:
modify cause items
Counterfactuality: backward vs forward

backward counterfactual causality

given an effect scenario:

“if the cause would not have happened, then the effect would not have occurred”

intervention:
modify cause items

forward counterfactual causality

given a world model:

“minimal set of items that need to be modified to avoid the effect”
Counterfactual causality: backward vs forward

backward counterfactual causality

given an effect scenario:
“if the cause would not have happened, then the effect would not have occurred”

intervention:
modify cause items

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
Given a transition system $MMM$ with state space $SSS$ and labeling functions $(L_s)_s \in S(L_s)$ where $L_s: AP \rightarrow \{0, 1\}$.

Intervention ("mutations of the truth values of atomic propositions"):\n
- Given $q \in AP$ and $T \subseteq ST \subseteq S$, then $M_{T, q}$ is $MMM$ with flipped labeling values $L_t(q)$ for $t \in T$.

Suppose $M \models \varphi$ and let $q \in AP$.

- Switching pair: $(T, s)$ where $T \subseteq ST \subseteq S$, $s \in S$ s.t. $M_{T, q} \models \varphi$ and $M_{T \cup \{s\}, q} \not\models \varphi$.

\[\text{state } s \text{ is a } q \text{-cause state for } M \models \varphi \text{ if there exists a switching pair } (T, s)\]

[Chockler/Halpern/Kupferman, ACM ToCL 2008]
Given a transition system $\mathcal{M}$ with state space $S$ and labeling functions $(L_s)_{s \in S}$ where $L_s : AP \rightarrow \{0, 1\}$.

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

[Chockler/Halpern/Kupferman, ACM ToCL 2008]
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 $t \in T$. 

[Chockler/Halpern/Kupferman, ACM ToCL 2008]
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 $t \in T$. 

Diagram:

- $\mathcal{M}$
- $T = \{t_1, t_2\}$
- $T = \{t_1, t_2\}$
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 $t \in T$. 

**Diagram:**

[Transition system $\mathcal{M}$ with states $s$, $t_1$, $t_2$, and transitions between them, labeled with propositions $p$ and $q$.]

[Transition system $\mathcal{M}_{T,q}$ with the same states and transitions, but $q$ is flipped in the labeling at $t_1$ and transitions to $t_2$.]
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 $t \in T$.

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

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.
  \[ \mathcal{M}_{T,q} \models \phi \quad \text{and} \quad \mathcal{M}_{T \cup \{s\},q} \not\models \phi \]
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 $t \in T$.

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 $L_s : AP \rightarrow \{0, 1\}$.

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 $t \in T$.

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

$AP = \{q\}$

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

$s_4 \models q$

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

$s_4$ 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

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

\[ M \models \exists q \]
\[ M_{T,q} \not\models \exists q \quad \text{iff} \quad T = \{ s_4 \} \]
Example: responsibility à la Chockler et al

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

\[ \mathcal{M} \models \exists \Diamond q \]
\[ \mathcal{M}_{T,q} \not\models \exists \Diamond q \quad \text{iff} \quad T = \{ s_4 \} \]

\( (\emptyset, s_4) \) is the only switching pair
Example: responsibility à la Chockler et al

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

\[ M \models \exists \diamond q \]
\[ M_{T,q} \not\models \exists \diamond q \text{ iff } T = \{ s_4 \} \]

\[ (\emptyset, s_4) \text{ is the only switching pair} \]

- \( s_4 \) 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

\[ \mathcal{M} \models \exists \Diamond q \]
\[ \mathcal{M}_{T,q} \not\models \exists \Diamond q \quad \text{iff} \quad T = \{s_3, s_4\} \]

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 \( M_T \) equals \( M \) 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

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

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

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

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

- **backward**: for a given execution where $\phi$ holds, which states were responsible for the satisfaction of $\phi$? 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 $\phi$ (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 $\phi$ (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 $\phi$)
Forward responsibility for temporal properties

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

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

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

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

\[ C \text{ is minimal w.r.t. } [F1] \]

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

Observations:

- If $\mathcal{M} |= \forall \phi$ then no one is forward responsible, and vice versa.
- If $\mathcal{M} |= \forall \neg \phi$ then exactly $C = \emptyset$ is forward responsible.
Forward responsibility for temporal properties

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

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

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

i.e., a strategy $\sigma$ for player 1 s.t. the bad event does not happen in $\sigma$-plays.
Forward responsibility for temporal properties

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

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

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

i.e., a strategy $\sigma$ for player 1 s.t. the bad event does not happen in $\sigma$-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 $\phi$ (bad event).

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

[F1] $C$ has a winning strategy in $\mathcal{M}_C$ for objective $\neg \phi$
    i.e., a strategy $\sigma$ for player 1 s.t. the bad event does not happen in $\sigma$-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 no one 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 $\phi$ (bad event).

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

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

i.e., a strategy $\sigma$ for player 1 s.t. the bad event does not happen in $\sigma$-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 $C = \emptyset$ is forward responsible.
Forward responsibility: example

\[ \phi = \lozenge \text{fail} \quad ("\text{bad event"}) \]

\(C\) is forward responsible for \(\phi\) if

[F1] \(C\) has a winning strategy in \(M_C\) for objective \(\neg\phi\)

[F2] \(C\) is minimal w.r.t. [F1]
**Forward responsibility: example**

\[\phi = \Diamond \text{fail} \quad (\text{"bad event"})\]

forward responsible sets:

\[\{t, u\}\]

**C** is forward responsible for **\(\phi\)** if

[F1] **C** has a winning strategy in \(M_C\) for objective \(\neg \phi\)

[F2] **C** is minimal w.r.t. [F1]
Forward responsibility: example

\[ \phi = \Box \text{fail} \quad (\text{"bad event"}) \]

forward responsible sets:

\[ \{t, u\} \]
\[ \{s, u\} \]

\( C \) is forward responsible for \( \phi \) if

[F1] \( C \) has a winning strategy in \( M_C \) for objective \( \neg \phi \)

[F2] \( C \) is minimal w.r.t. [F1]
Forward responsibility: example

Goal, Goal, ϕ = ♢ fail

forward responsible sets:

{t, u}
{s, u}
{s, t}

ϕ = ♢ 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]
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

Given $\mathbb{M}$, path property $\phi$, a set $\mathbb{C}$ of states and a path $\pi = s_0 s_1 s_2 \ldots \pi = s_0 s_1 s_2 \ldots$ s.t. $\pi|_i = \phi \pi|_i = \phi \pi|_i = \phi$.

$\mathbb{C}$ is strategically backward responsible for "$\pi|_i = \phi \pi|_i = \phi \pi|_i = \phi$" if there exists $n \in \mathbb{N}$ such that $\mathbb{C}$ has a winning strategy in $\mathbb{M}$ for objective $\neg \phi$ from state $s_n i.e., \mathbb{C}$ could have played differently from $s_n$ to enforce the violation of $\phi$. $\mathbb{C}$ is minimal w.r.t. $[SB1]$.
Strategic backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 there exists $n \in \mathbb{N}$ such that $C$ has a winning strategy in $M$ for objective $\neg \phi$ from state $s_n$, i.e., $C$ could have played differently from $s_n$ to enforce the violation of $\phi$. $C$ is minimal w.r.t. [SB1].
Strategic backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\phi$
Strategic backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\phi$

[SB2] $C$ is minimal w.r.t. [SB1]
Strategic backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\phi$

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

objective from state $s_n$: $\neg \phi$ if $\phi$ 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 $\phi$, a set $C$ of states and a deterministic strategy profile $\sigma = (\sigma_C, \sigma_{\overline{C}})$, $\mathcal{C}$ is causally backward responsible for "$\mathcal{M}, \sigma \models \phi$" if

1. there exists a strategy $\tau_C$ for $\mathcal{C}$ in $\mathcal{M}$ such that the unique $(\tau_C, \sigma_C)$-play satisfies $\neg \phi$,
   i.e., $\mathcal{C}$ could have played differently to enforce the violation of $\phi$, when the strategy for the other states is fixed

2. $\mathcal{C}$ is minimal w.r.t. (CB1) i.e., no proper subset of $\mathcal{C}$ can enforce the violation of $\phi$, when the other states stick to their strategy
Causal backward responsibility

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

Strategy profile $\sigma$ specifies
- a path (the unique $\sigma$-play $\pi_{\sigma}$)
- $\overline{\mathcal{C}}$’s decision along other paths (for counterfactual reasoning)
- $\mathcal{C}$’s decision along other paths (irrelevant)
Causal backward responsibility

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

\[ \pi_\sigma \models \phi \]
for the unique $\sigma$-play $\pi_\sigma$

Strategy profile $\sigma$ specifies
- a path (the unique $\sigma$-play $\pi_\sigma$)
- $\overline{\mathcal{C}}$’s decision along other paths (for counterfactual reasoning)
- $\mathcal{C}$’s decision along other paths (irrelevant)
Causal backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\tau_C$ for $C$ in $\mathcal{M}_C$ s.t. the
unique $(\tau_C, \sigma_{\overline{C}})$-play satisfies $\neg \phi$

Strategy profile $\sigma$ specifies

- a path (the unique $\sigma$-play $\pi_\sigma$)
- $\overline{C}$’s decision along other paths (for counterfactual reasoning)
- $C$’s decision along other paths (irrelevant)
Causal backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\tau_C$ 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 $\phi$, when the strategy for the other states is fixed.
Causal backward responsibility

Given TS $\mathcal{M}$, path property $\phi$, 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 $\tau_C$ 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 $\phi$, 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 $\phi$, when the other states stick to their strategy
Backward responsibility: example (strategic)

\[ \phi = \Diamond \text{fail} \] ("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)

\[ \phi = \Diamond \text{fail} \] ("bad event")

\[
\text{path } s \rightarrow t \rightarrow \text{fail} \models \phi
\]

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)

\[ \phi = \Diamond \text{fail} \quad ("bad \ event") \]

path \( s \ t \ \text{fail} \models \phi \)

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)

\[ \phi = \Diamond \text{fail} \text{ ("bad event") } \]

path \( s \to t \to \text{fail} \models \phi \)

strat-backward responsible:

\[ \{s, u\} \]
\[ \{t\} \]

\[ C \text{ is strategically backward responsible for } s_0 s_1 s_2 \ldots \models \phi \text{ if } \]

[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)

\[ \phi = \Diamond fail \text{ ("bad event")} \]

\[
\begin{align*}
C \text{ is causally backward responsible for } (\sigma_C, \sigma_{\neg C}) & \models \phi \text{ if } \\
[CB1] & \text{ there is a strategy } \tau_C \text{ for } C \text{ s.t. the } (\tau_C, \sigma_{\neg C})\text{-play satisfies } \neg \phi \\
[CB2] & C \text{ is minimal w.r.t. [CB1]} 
\end{align*}
\]
Backward responsibility: example (causal)

\[ \phi = \lozenge \text{fail} \] ("bad event")

strategy profile:
\[ s \rightarrow t, \ t \rightarrow f, \ u \rightarrow g_2 \]

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

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

[CB2] C is minimal w.r.t. [CB1]
Backward responsibility: example (causal)

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

strategy profile:

\[
\begin{align*}
\tau C: & s \rightarrow t, \quad t \rightarrow f, \quad u \rightarrow g_2
\end{align*}
\]

causally backward responsible:

\[
\{ t \}; \text{ change } t \rightarrow g_1
\]

\( C \) is causally backward responsible for \((\sigma C, \sigma \bar{C}) \models \phi \) if

[CB1] there is a strategy \(\tau C\) for \( C \) s.t. the \((\tau C, \sigma \bar{C})\)-play satisfies \( \neg \phi \)

[CB2] \( C \) is minimal w.r.t. [CB1]
Backward responsibility: example (causal)

\( \phi = \Box \text{fail} \) ("bad event")

strategy profile:
\[
\begin{align*}
  s &\to t, \\
  t &\to f, \\
  u &\to g_2
\end{align*}
\]

causally backward responsible:
\[
\{t\}; \text{ change } t \to g_1 \\
\{s\}; \text{ change } s \to u
\]

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

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

[CB2] \( C \) is minimal w.r.t. [CB1]
Relation between f-, sb- and cb-responsibility

\[ f \text{-responsible} \Rightarrow sb \text{-responsible} \Rightarrow cb \text{-responsible} \]

Let \( CCC \) be a set of states.

- \( CCC \) is f-responsible for \( \phi \) if \( CCC \) contains a coalition that is sb-responsible for all \( \pi |\pi| = \phi \), and is minimal w.r.t. this property.
- If \( CCC \) is sb-responsible for \( \pi |\pi| = \phi \) and \( \sigma \) a strategy profile s.t. \( \pi \) is the \( \sigma \)-play then \( CCC \) contains a coalition that is cb-responsible for \( M |\sigma| = \phi \), \( M |\sigma| = \phi \), \( M |\sigma| = \phi \).

f-responsible = forward responsible
sb-responsible = strategically backward responsible
cb-responsible = causally backward responsible
Relation between f-, sb- and cb-responsibility

\[
\text{f-responsibility} \implies \text{sb-responsibility} \implies \text{cb-responsibility}
\]

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

Let $C$ be a set of states.

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

f-responsible $\implies$ sb-responsibility $\implies$ cb-responsibility
Relation between f-, sb- and cb-responsibility

\[ \text{f-responsibility} \implies \text{sb-responsibility} \implies \text{cb-responsibility} \]

Let \( C \) be a set of states.

- \( C \) is f-responsible for \( \phi \) 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 \( \sigma \) a strategy profile s.t. \( \pi \) is the \( \sigma \)-play then \( C \) contains a coalition that is cb-responsible for \( M, \sigma \models \phi \).
Relation between f-, sb- and cb-responsibility

Let $C$ be a set of states.

- $C$ is f-responsible for $\phi$ 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 $\sigma$ a strategy profile s.t. $\pi$ is the $\sigma$-play then $C$ contains a coalition that is cb-responsible for $M, \sigma \models \phi$.

generalizes HP-causality in SEM
HP-causality and cb-responsibility

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

$\Downarrow$

tree-like transition system $\mathcal{M}_{S,c}$
HP-causality and cb-responsibility

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

\[ \downarrow \]

total order for endo variables: \( x_1, \ldots, x_n \)

tree-like transition system \( M_{S,c} \)

- 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: \( s \rightarrow [s, x_i=f_i(c, s)] \)
  - intervention: \( s \rightarrow [s, x_i=\beta] \) for any other value \( \beta \)
HP-causality and cb-responsibility

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

\[ \downarrow \]

tree-like transition system \( M_{S,c} \)

Given a Boolean condition \( \varphi \) for the endogenous variables:

\[ X=\alpha \text{ is a but-for cause for } \varphi \]
iff the \( X \)-states constitute a cb-responsible coalition for \( \phi \) under the default strategy profile

where \( \phi = \Diamond \text{“}\varphi \text{ holds at some leave”} \) and \( \alpha = S_X(c) \)
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

Lloyd S. Shapley
(Nobel prize 2012 for Economics)
Cooperative games and Shapley values

Cooperative game: one-shot game consisting of

- a finite set of agents, say $A_g = \{1, \ldots, n\}$
- a payoff function $\text{val}: 2^{A_g} \rightarrow \mathbb{R}$
  such that $\text{val}(\emptyset) = 0$

Given a total order $\pi \subseteq A_g$ and an agent $a \in A_g$:

$$\text{Sh}(a) = \frac{1}{n!} \sum_{\pi \in \Pi_n} \left( \text{val}(\pi \uparrow a) - \text{val}(\pi^\downarrow a) \right)$$

The Shapley value is an average contribution of agent $a$ to the value of coalition $\pi \uparrow a$.

The average contribution of agent $a$ is:

$$\frac{1}{C \subseteq A_g \mid a \in C} \left( n - |C| - 1 \right)! n! \text{val}(C \cup \{a\}) - \text{val}(C)$$
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. $val(\emptyset) = 0$

$$val(C) = \text{value of coalition } C \subseteq Ag$$
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 $val : 2^{Ag} \rightarrow \mathbb{R}$ s.t. $val(\emptyset) = 0$

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

$$\pi \succeq a = \{ i \in Ag \mid \pi(i) \geq \pi(a) \}$$
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 $val : 2^{Ag} \rightarrow \mathbb{R}$ s.t. $val(\emptyset) = 0$

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

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

$$val(\pi \geq a) - val(\pi > a)$$

contribution of agent $a$ to the value of coalition $\pi \geq a$
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. $val(\emptyset) = 0$

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

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

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

contribution of agent $a$ to the value of coalition $\pi \geq a$

“average contribution of agent $a$”
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. $val(\emptyset) = 0$

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

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

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

$$= \sum_{\substack{C \subseteq Ag \\setminus a \\subseteq C}} \frac{|C|!(n-|C|-1)!}{n!} (val(C \cup \{a\}) - val(C))$$
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 $\phi$ (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 $\phi$ (e.g. LTL formula).

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

[Mascl/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 $\phi$ (e.g. LTL formula).

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

Game-based view:

- states may build coalitions that attempt to enforce $\phi$ 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 $\phi$ 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 $\phi$ (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 \( \phi \) (e.g. LTL formula).
Let \( C \subseteq S \) and \( \mathcal{M}_C \) as before with objective \( \phi \) 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 \( \phi \))
- player 2 controls all states in \( \bar{C} = S \setminus C \) (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 $\phi$ (e.g. LTL formula).

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

Payoff value of coalition $C$:

$$\text{val}_{\phi}(C) = \begin{cases} 
1 : & \text{if } C \text{ has a winning strategy in } \mathcal{M}_C \text{ for } \phi \\
0 : & \text{otherwise}
\end{cases}$$
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 $\phi$ (e.g. LTL formula).

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

Payoff value of coalition $C$:

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

Importance value of state $s = \text{Shapley value of } s$ in the simple cooperative game with agent set $Ag = S$ and payoff function $\text{val}_\phi$. 
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 $\phi$ (e.g. LTL formula).

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

Payoff value of coalition $C$:

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

Importance value of state $s = \text{Shapley value of } s$

in the simple cooperative game with agent set $Ag = S$ and payoff function $\text{val}_\phi$

0/1-values and monotonicity, i.e., if $C \subseteq D$ then $\text{val}_\phi(C) \leq \text{val}_\phi(D)$
Importance values: properties

Importance value of state $s = \text{Shapley value of } s$

$$I_\phi(s) = \sum_{\substack{C \subseteq S \setminus \{s\} \subseteq C \subseteq S \setminus \{s\}}} \frac{|C|!(n-|C|-1)!}{n!} \left( \text{val}_\phi(C \cup \{s\}) - \text{val}_\phi(C) \right)$$

$n = |S|$
Importance values: properties

Importance value of state $s = \text{Shapley value of } s$

$$I_\phi(s) = \sum_{\substack{C \subseteq S \setminus \{s\} \subseteq S \setminus \{s\} \subseteq S \setminus \{s\} = \text{set of relevant states} \quad \text{or } 0 \text{ or } 1}} \frac{|C|!(n-|C|-1)!}{n!} \left( \text{val}_\phi(C \cup \{s\}) - \text{val}_\phi(C) \right)$$

where $(C, s)$ is switching iff $\text{val}_\phi(C \cup \{s\}) = 1$ and $\text{val}_\phi(C) = 0$
Importance values: properties

Importance value of state $s = \text{Shapley value of } s$

\[
\mathcal{I}_\phi(s) = \sum_{\substack{C \subseteq S \subseteq R \subseteq S \ni s \neq C}} \frac{|C|!(n-|C|-1)!}{n!} \left( \text{val}_\phi(C \cup \{s\}) - \text{val}_\phi(C) \right)
\]

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

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

Importance value of state $s = \text{Shapley value of } s$

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

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

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

$I_\phi(s) > 0$ iff $s$ is relevant, i.e., there is a switching pair $(C, s)$

A switching pair $(C, s)$ is relevant iff $C \subseteq R = \text{set of relevant states}$
Importance values: properties

Importance value of state \( s \) = Shapley value of \( s \)

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

where \( r = \# \) 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 $s = \text{Shapley value of } s$

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

where $r = \# \text{ relevant states}$

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

$$\text{val}_\phi(C) = 1 - \text{val}_{\neg \phi}(\overline{C})$$

$(C, s)$ relevant for $\phi$ iff $(\overline{C} \cap R \setminus \{s\}, s)$ relevant for $\neg \phi$
Importance values: properties

Importance value of state \( s \) = Shapley value of \( s \)

\[
I_{\phi}(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!}
\]

where \( r \) = \# relevant states

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

\[
\text{val}_\phi(C) = 1 - \text{val}_{\neg\phi}(\overline{C})
\]

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

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

where \( r = \# \text{ relevant states} \)

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

\[
\text{val}_\phi(C) = 1 - \text{val}_{\neg\phi}(\overline{C})
\]

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

Hence:

\[
\mathcal{I}_\phi(s) = \mathcal{I}_{\neg\phi}(s)
\]

“importance of states on the truth value (satisfaction or violation) of \( \phi \)"
Importance values: example

\[ \phi = \Box \lozenge s \land \lozenge \Box \neg f \]

Importance value of state \( s \) = Shapley value of \( s \)

\[ I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|! (r - |C| - 1)!}{r!} \]

where \( r = \# \) relevant states
Importance values: example

\[ \phi = \Box \lozenge s \land \lozenge \Box \neg f \]

deterministic states are irrelevant

(Importance value 0)

Importance value of state \( s = \) Shapley value of \( s \)

\[ I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!} \]

where \( r = \# \text{ relevant states} \)
Importance values: example

\[ \phi = \square \Diamond s \land \Diamond \square \neg f \]

deterministic states are irrelevant (importance value 0)

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

Importance value of state \( s = \) Shapley value of \( s \)

\[ I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!} \]

where \( r = \# \) relevant states
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

deterministic states are irrelevant

(importance value 0)

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

\[
I_\phi(w) = I_\phi(g) = \frac{1!(2-1-1)!}{2!} = \frac{1!0!}{2!} = \frac{1}{2}
\]

Importance value of state \(s = \text{Shapley value of } s\)

\[
I_\phi(s) = \sum_{(C, s), \text{relevant}} \frac{|C|!(r-|C|-1)!}{r!}
\]

where \(r = \# \text{ relevant states}\)
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

Importance value of state \( s = \) Shapley value of \( s \)

\[ I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!} \]

where \( r = \# \text{ relevant states} \)
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

state \( f \) is irrelevant

Importance value of state \( s \) = Shapley value of \( s \)

\[
I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!}
\]

where \( r = \# \) relevant states
Importance values: example

\[ \phi = \Box \lozenge s \land \lozenge \Box \neg f \]

state \( f \) is irrelevant

\( C \) has a winning strategy iff

\( g \in C \) and \(|C \cap \{w_1, w_2, s\}| \geq 2\)

Importance value of state \( s \) = Shapley value of \( s \)

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

where \( r = \# \) relevant states
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

state \( f \) is irrelevant

\( C \) has a winning strategy iff \( g \in C \) and \( |C \cap \{w_1, w_2, s\}| \geq 2 \)

In particular: \( r = 4 \)

Importance value of state \( s = \) Shapley value of \( s \)

\[ I_\phi(s) = \sum_{(C, s) \text{ relevant}} \frac{|C|!(r-|C|-1)!}{r!} \]

where \( r = \# \) relevant states
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

State \( f \) is irrelevant

\( C \) has a winning strategy iff \( g \in C \) and \( |C \cap \{w_1, w_2, s\}| \geq 2 \)

In particular: \( r = 4 \)

4 relevant pairs for \( g \) and \[ I_\phi(g) = 3 \cdot \frac{2!(4-2-1)!}{4!} + \frac{3!(4-3-1)!}{4!} = \frac{1}{2} \]
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

state \( f \) is irrelevant

\( C \) has a winning strategy iff

\( g \in C \) and \( |C \cap \{w_1, w_2, s\}| \geq 2 \)

In particular: \( r = 4 \)

4 relevant pairs for \( g \) and
\[ I_\phi(g) = 3 \cdot \frac{2!(4-2-1)!}{4!} + \frac{3!(4-3-1)!}{4!} = \frac{1}{2} \]

2 relevant pairs for \( w_1 \) and
\[ I_\phi(w_1) = 2 \cdot \frac{2!(4-2-1)!}{4!} = \frac{1}{6} \]
Importance values: example

\[ \phi = \Box \Diamond s \land \Diamond \Box \neg f \]

State \( f \) is irrelevant

\( C \) has a winning strategy iff \( g \in C \) and \( |C \cap \{w_1, w_2, s\}| \geq 2 \)

In particular: \( r = 4 \)

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

2 relevant pairs for \( w_1 \) and \( I_\phi(w_1) = 2 \cdot \frac{2!(4-2-1)!}{4!} = \frac{1}{6} \)
Importance values: algorithmic problems

For transition system $M$ with state space $S$ and path property $\phi$.

Value problem:
  given $C \subseteq S$, check whether $\text{val}_\phi(C) = 1$

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 $M$ with state space $S$ and path property $\phi$.

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

Usefulness problem: given state $s$, decide whether $I_\phi(s) > 0$

Importance problem: given state $s$, compute $n! I_\phi(s)$
Importance values: algorithmic problems

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

Value problem: ... standard game solving
given $C \subseteq S$, check whether $val_{\phi}(C) = 1$

Usefulness problem:
given state $s$, decide whether $I_{\phi}(s) > 0$

Importance problem:
given state $s$, compute $n! I_{\phi}(s)$

Solving the usefulness and importance problems, via standard game solving algorithms + guessing relevant pairs.
## Importance values: complexity results

<table>
<thead>
<tr>
<th></th>
<th>Büchi</th>
<th>Rabin</th>
<th>Streett</th>
<th>Parity</th>
<th>LTL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value problem</strong></td>
<td>$P$</td>
<td>NP</td>
<td>coNP</td>
<td>$\in \text{NP} \cap \text{coNP}$</td>
<td>$2\exp$</td>
</tr>
<tr>
<td><strong>Usefulness problem</strong></td>
<td>NP</td>
<td>$\Sigma_2^P$</td>
<td>$\Sigma_2^P$</td>
<td>NP</td>
<td>$2\exp$</td>
</tr>
<tr>
<td><strong>Importance problem</strong></td>
<td>$#P$</td>
<td>$#P^{NP}$</td>
<td>$#P^{NP}$</td>
<td>$#P$</td>
<td>$2\exp$</td>
</tr>
</tbody>
</table>
## Importance values: complexity results

<table>
<thead>
<tr>
<th></th>
<th>Büchi</th>
<th>Rabin</th>
<th>Streett</th>
<th>Parity</th>
<th>LTL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value problem</strong></td>
<td>P</td>
<td>NP</td>
<td>coNP</td>
<td>∈ NP ∩ coNP</td>
<td>2EXP</td>
</tr>
<tr>
<td><strong>Usefulness problem</strong></td>
<td>NP</td>
<td>Σ₂^P</td>
<td>Σ₂^P</td>
<td>NP</td>
<td>2EXP</td>
</tr>
<tr>
<td><strong>Importance problem</strong></td>
<td>#P</td>
<td>#P^{NP}</td>
<td>#P^{NP}</td>
<td>#P</td>
<td>2EXP</td>
</tr>
</tbody>
</table>

Value problem: classical results for games
## Importance values: complexity results

<table>
<thead>
<tr>
<th>Value problem</th>
<th>Büchi</th>
<th>Rabin</th>
<th>Streett</th>
<th>Parity</th>
<th>LTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>NP</td>
<td>coNP</td>
<td>∈ NP ∩ coNP</td>
<td>2EXP</td>
<td></td>
</tr>
</tbody>
</table>

### Usefulness problem

| NP
| \( \Sigma_2^P \) | \( \Sigma_2^P \) | NP | 2EXP |

### Importance problem

| \#P | \#P_{NP} | \#P_{NP} | \#P | 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

<table>
<thead>
<tr>
<th></th>
<th>Büchi</th>
<th>Rabin</th>
<th>Streett</th>
<th>Parity</th>
<th>LTL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value problem</strong></td>
<td>P</td>
<td>NP</td>
<td>coNP</td>
<td>∈ NP ∩ coNP</td>
<td>2EXP</td>
</tr>
<tr>
<td><strong>Usefulness problem</strong></td>
<td>NP</td>
<td>$\Sigma_2^P$</td>
<td>$\Sigma_2^P$</td>
<td>NP</td>
<td>2EXP</td>
</tr>
<tr>
<td><strong>Importance problem</strong></td>
<td>#P</td>
<td>#P$^{NP}$</td>
<td>#P$^{NP}$</td>
<td>#P</td>
<td>2EXP</td>
</tr>
</tbody>
</table>

$\Sigma_2^P$-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)
- $\Sigma_2^P$-hardness via reduction from dual of $\forall\exists 3$SAT
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
Probabilistic causality

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

Reichenbach (1956)
Suppes (1970)
and many more

Judea Pearl
Turing Award
Winner 2011
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:**
  \[
  \Pr(\text{effect} | \text{cause}) > \Pr(\text{effect} | \neg\text{cause})
  \]
Probabilistic causality

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

Two main principles:

**Temporal condition:**

Causes occur before their effects.

**Probability-raising condition:**

\[
\Pr(\text{effect} | \text{cause}) > \Pr(\text{effect} | \neg \text{cause})
\]

equivalently: \[
\Pr(\text{effect} | \text{cause}) > \Pr(\text{effect})
\]
Probabilistic causality

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

Two main principles:

**Temporal condition:**
Causes occur before their effects.

**Probability-raising condition:**
\[ \Pr(\text{effect} | \text{cause}) > \Pr(\text{effect} | \neg \text{cause}) \]

Probabilistic form of counterfactuality:
“effects are less likely if their causes do not occur”
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
  [´Abrah´am/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
  [´Abrah´am/Bonakdarpour, QEST'18]

- cause-effect relations for regular causes and ω-regular effects in
  Markov chains
  [Dimitrova/Finkbeiner/Torfah, ATVA'20]

- cause-effect relations for sets of states in Markov decision processes
  [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]

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 $\omega$-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/Torfa, ATVA’20]

- cause-effect relations for regular causes and $\omega$-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 $\omega$-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]
In what follows:

- \( \mathcal{M} \) is a (discrete-time) Markov chain with
  - finite state space \( S \)
  - initial distribution \( \pi : S \to [0, 1] \) such that every state in \( S \) is accessible from at least one initial state (i.e., a state \( s \) with \( \pi(s) > 0 \))
  - a fixed nonempty set \( E \) of effect states
    - W.l.o.g. all \( E \) states are terminal (i.e., do not have outgoing transitions).

The effect probability in \( \mathcal{M} \) is:

\[
\Pr_M(\diamond E) \triangleq \sum_{s \in S} \pi(s) \cdot \Pr_s(\diamond E)
\]

The effect probability from state \( s \) is:

\[
\Pr_s(\diamond E)
\]
Probabilistic causality in Markov chains

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

- finite state space $S$
- initial distribution $\nu : S \to [0, 1]$ such that every state in $S$ is accessible from at least one initial state (i.e., a state $s$ with $\nu(s) > 0$)
- 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 $\nu : S \rightarrow [0, 1]$ such that every state in $S$ is accessible from at least one initial state (i.e., a state $s$ with $\nu(s) > 0$)
- 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).

$\Pr_{\mathcal{M}}(\diamondsuit E)$ effect probability in $\mathcal{M}$

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

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

\[
\text{Pr}_\mathcal{M}(\diamond E) \quad \text{effect probability in } \mathcal{M} \quad = \sum_{s \in S} \nu(s) \cdot \text{Pr}_s(\diamond E)
\]

\[
\text{Pr}_s(\diamond E) \quad \text{effect probability from state } s
\]
PCTL-characterization of causality in MC

Let $C$ be a set of states with $C \cap E = \emptyset$. $C$ is called a (prima facie) cause for $E$ if there exists $p \in (0, 1)$ such that $M |\triangledown E < p$ and $M |\forall \Box C \rightarrow P \geq p(\triangledown E)$ for all $s \in C$.

[PCTL: probabilistic computation logic]

[Kleinberg, PhD thesis 2010]
PCTL-characterization of causality in MC

Let $C$ a set of states with $C \cap E = \emptyset$. $C$ is called a \textit{(prima facie) cause} for $E$ if there exists $p \in ]0, 1]$ s.t.

$$M \models \mathbb{P}_p(\Diamond E) \quad \text{and} \quad M \models \forall \Box (C \rightarrow \mathbb{P}_{\geq p}(\Diamond E))$$

[Kleinberg, PhD thesis 2010]
Let $C$ a set of states with $C \cap E = \emptyset$. 

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

\[
M \models \Pr_{<p}(\Diamond E) \quad \text{and} \quad M \models \forall \Box (C \rightarrow \Pr_{\geq p}(\Diamond E))
\]

\[
\text{Pr}_M(\Diamond E) < p
\]

[Kleinberg, PhD thesis 2010]
PCTL-characterization of causality in MC

Let $C$ a set of states with $C \cap E = \emptyset$. $C$ is called a (prima facie) cause for $E$ if there exists $p \in ]0, 1]$ s.t.

\[
\mathcal{M} \models \mathbb{P}_{<p}(\diamond E) \quad \text{and} \quad \mathcal{M} \models \forall \Box (C \rightarrow \mathbb{P}_{\geq p}(\diamond E))
\]

\[
\Pr_{\mathcal{M}}(\diamond E) < p \quad \text{Pr}_{s}(\diamond E) \geq p
\]

for all $s \in C$

[Kleinberg, PhD thesis 2010]
PCTL-characterization of causality in MC

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

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

\[
\mathcal{M} \models P_{<p}(\Diamond E) \quad \text{and} \quad \mathcal{M} \models \forall \Box (C \rightarrow P_{\geq p}(\Diamond E))
\]

Thus:

\[
\Pr_{\mathcal{M}}(\Diamond E) < p \quad \text{for all } s \in C
\]

\[
\Pr_{s}(\Diamond E) \geq p
\]

Thus:

$C$ cause for $E$ iff $\Pr_{\mathcal{M}}(\Diamond E) < \Pr_{s}(\Diamond E)$ for all $s \in C$
Let \( C \) a set of states with \( C \cap E = \emptyset \).

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

\[
\mathcal{M} \models \mathbb{P}_{<p}(\Diamond E) \quad \text{and} \quad \mathcal{M} \models \forall \Box (C \rightarrow \mathbb{P}_{\geq p}(\Diamond E))
\]

\[
\Pr_{\mathcal{M}}(\Diamond E) < p \quad \text{and} \quad \Pr_{s}(\Diamond E) \geq p
\]

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 \( s \in C \)
PCTL-characterization of causality in MC

Let \( C \) a set of states with \( C \cap E = \emptyset \).

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

\[
\mathcal{M} \models \mathbb{P}_p(\Diamond E) \quad \text{and} \quad \mathcal{M} \models \forall \Box (C \rightarrow \mathbb{P}_p(\Diamond E))
\]

\[
\Pr_{\mathcal{M}}(\Diamond E) < p \quad \text{and} \quad \Pr_s(\Diamond E) \geq p
\]

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 $C \cap E = \emptyset$.

- $C$ is a strict probability-raising (SPR) cause for $E$ iff
  \[
  \Pr_M(\diamondsuit E) < \Pr_M(\diamondsuit E | \diamondsuit s) \quad \text{for all } s \in C
  \]

Each SPR cause is a GPR cause.

If $C$ is a singleton then:
- $C$ is a SPR cause iff $C$ is a GPR cause
Strict/global probability-raising causes in MC

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

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

\[
\Pr_M(\Box E) < \Pr_M(\Box E | \Box s) \quad \text{for all } s \in C
\]

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

\[
\Pr_M(\Box E) < \Pr_M(\Box E | \Box C)
\]

Each SPR cause is a GPR cause.

If $C$ is a singleton then:

$C$ is a SPR cause iff $C$ is a GPR cause.
Strict/global probability-raising causes in MC

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

- $C$ is a strict probability-raising (SPR) cause for $E$ iff
  $$\Pr_M(\Diamond E) < \Pr_M(\Diamond E | \Diamond s)$$
  for all $s \in C$.

- $C$ is a global probability-raising (GPR) cause for $E$ iff
  $$\Pr_M(\Diamond E) < \Pr_M(\Diamond E | \Diamond C)$$
  plus 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 $\pi$ in $M$ with $\pi \models (\neg C) \cup s$. 

Strict/global probability-raising causes in MC

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

- $C$ is a strict probability-raising (SPR) cause for $E$ iff
  $$\Pr_M(\diamondsuit E) < \Pr_M(\diamondsuit E | \diamondsuit s)$$
  for all $s \in C$

- $C$ is a global probability-raising (GPR) cause for $E$ iff
  $$\Pr_M(\diamondsuit E) < \Pr_M(\diamondsuit E | \diamondsuit C)$$
  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 $C \cap E = \emptyset$.

- $C$ is a strict probability-raising (SPR) cause for $E$ iff
  \[ \Pr_M(\diamond E) < \Pr_M(\diamond E | \diamond s) \] for all $s \in C$

- $C$ is a global probability-raising (GPR) cause for $E$ iff
  \[ \Pr_M(\diamond E) < \Pr_M(\diamond E | \diamond C) \]
  plus some minimality constraint (omitted here)

- Each SPR cause is a GPR cause.

- If $C$ is a singleton then:
  \[ C \text{ is a SPR cause } \iff \text{C is a GPR cause} \]
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 $E = \{e_1, e_2\}$

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

$C = \{c_1, c_2\}$

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

- $C$ is a GPR cause as

$$\Pr_{\mathcal{M}}(\diamondsuit E | \diamondsuit C) = \frac{\frac{1}{3} + \frac{1}{3} \cdot \frac{1}{4}}{\frac{1}{3} + \frac{1}{3}} = \frac{\frac{5}{12}}{\frac{2}{3}} = \frac{5}{8} > \frac{1}{2} = \Pr_{\mathcal{M}}(\diamondsuit E)$$
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 $E = \{e_1, e_2\}$

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

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

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

$\Pr_{\mathcal{M}}(\lozenge E | \lozenge C) = \frac{1}{2} = \Pr_{\mathcal{M}}(\lozenge E)$
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 $C \subseteq \{s_1, s_2, s_3\}$:

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

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

[Diagram of MC with states and transitions]
Markov decision processes (MDP)

... extension of Markov chains by nondeterministic choices ...
Markov decision processes (MDP)

- finite state space $S$ with initial distribution $\nu : S \rightarrow [0, 1]$
- finite set of action $\text{Act}$
- for each state $s \in S$:
  - $\text{Act}(s)$: set of enabled actions in state $s$
  - for each action $\alpha \in \text{Act}(s)$: distribution $P_{s,\alpha} : S \rightarrow [0, 1]$ for the $\alpha$-successors of $s$
Markov decision processes (MDP)

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

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

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

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

- GPR cause for $E$ iff

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

**SPR:** strict probability-raising

**GPR:** global probability-raising
PR causes in MCs (repetition)

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

- SPR cause for $E$ iff for all $s \in C$
  \[ \Pr_M(\Diamond E) < \Pr_M(\Diamond E | (\neg C) \cup s) \]

- GPR cause for $E$ iff
  \[ \Pr_M(\Diamond E) < \Pr_M(\Diamond E | \Diamond C) \]

SPR: strict probability-raising
GPR: global probability-raising
PR causes in MDPs

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

- SPR cause for $E$ iff for all $s \in C$ and all schedulers $\sigma$:
  $$\Pr_{\mathcal{M}}^\sigma(\Diamond E) < \Pr_{\mathcal{M}}^\sigma(\Diamond E \mid \neg C \cup s)$$

- GPR cause for $E$ iff for all schedulers $\sigma$:
  $$\Pr_{\mathcal{M}}^\sigma(\Diamond E) < \Pr_{\mathcal{M}}^\sigma(\Diamond E \mid \Diamond C)$$

$$\Pr_{\mathcal{M}}^\sigma(...) = \begin{cases} \text{probability measure of the Markov chain} \\ \text{induced by scheduler } \sigma \end{cases}$$
PR causes in MDPs

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

- SPR cause for $E$ iff for all $s \in C$ and all schedulers $\sigma$:
  $$\Pr_{\mathcal{M}}^\sigma(\diamond E) < \Pr_{\mathcal{M}}^\sigma(\diamond E \mid (\neg C) \cup s) \text{ if } \Pr_{\mathcal{M}}^\sigma((\neg C) \cup s) > 0$$

- GPR cause for $E$ iff for all schedulers $\sigma$:
  $$\Pr_{\mathcal{M}}^\sigma(\diamond E) < \Pr_{\mathcal{M}}^\sigma(\diamond E \mid \diamond C) \text{ if } \Pr_{\mathcal{M}}^\sigma(\diamond C) > 0$$

$$\Pr_{\mathcal{M}}^\sigma(\ldots) = \begin{cases} \text{probability measure of the Markov chain} \\ \text{induced by scheduler } \sigma \end{cases}$$
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 $i$
eq

effect set $E = \{e\}$
eq

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

Consequence:

Memory can be needed for refuting the PR condition!
Example: PR cause in MDP

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

effect set $E = \{e\}$

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

No
Example: PR cause in MDP

Consider the scheduler $\sigma$ that schedules $\beta$ for the first visit of $s$ and $\alpha$ for the second visit of $s$.

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

effect set $E = \{e\}$

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

No

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 $E = \{e\}$

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

No

Consider the scheduler $\sigma$ that schedules $\beta$ for the first visit of $s$ and $\alpha$ for the second visit of $s$.

\[
\Pr^\sigma_{\mathcal{M}}(\diamond E) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{16}
\]
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 $\sigma$ that schedules $\beta$ for the first visit of $s$ and $\alpha$ for the second visit of $s$.

$$\Pr_\mathcal{M}^{\sigma}(\Diamond E) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{16} > \frac{1}{4} = \Pr_\mathcal{M}^{\sigma}(\Diamond E | \Diamond c)$$
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 the scheduler $\sigma$ that schedules $\beta$ for the first visit of $s$ and $\alpha$ for the second visit of $s$.

$$\Pr_{\mathcal{M}}^{\sigma}(\Diamond E) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{16} > \frac{1}{4} = \Pr_{\mathcal{M}}^{\sigma}(\Diamond E | \Diamond c)$$
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

Consider 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 $\sigma(s)(\beta) = 1 - \lambda$.

$$\Pr^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \Pr^\sigma_s(\Diamond E) < \Pr^\sigma_s(\Diamond E)$$

Some positive value
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 $\sigma(s)(\beta) = 1 - \lambda$.

$$\Pr^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \Pr^\sigma_s(\Diamond E) < \Pr^\sigma_s(\Diamond E) = \Pr^\sigma_c(\Diamond E)$$

some positive value

Consequence: Memory can be needed for refuting the PR condition!
**Example: PR cause in MDP**

Consider MR-scheduler $\sigma = \sigma_\lambda$ with $\sigma(s)(\alpha) = \lambda$ and $\sigma(s)(\beta) = 1 - \lambda$.

$$\Pr^\sigma_{\mathcal{M}}(\lozenge E) = \frac{1}{2} \cdot \Pr^\sigma_s(\lozenge E) < \Pr^\sigma_s(\lozenge E) = \Pr^\sigma_c(\lozenge E) = \Pr^\sigma_{\mathcal{M}}(\lozenge E | \lozenge c)$$
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 $\sigma(s)(\beta) = 1 - \lambda$.

$$\Pr^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \Pr^\sigma_s(\Diamond E) < \Pr^\sigma_s(\Diamond E) = \Pr^\sigma_c(\Diamond E) = \Pr^\sigma_{\mathcal{M}}(\Diamond E | \Diamond c)$$

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?

---

**Diagram:**
- Initial state $i$ with transitions:
  - $\alpha, 1$ to $e_2$
  - $\beta, \frac{1}{2}$ to $\emptyset$
  - $\beta, \frac{1}{2}$ to $c$
  - $\frac{1}{2}$ to $e_2$
  - $\frac{1}{2}$ to $c$

---
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
Example: PR cause in MDP

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

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

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

No

Consider the scheduler $\sigma$ that schedules $\alpha$ and $\beta$ with probability $1/2$ in state $i$. 

\[
\begin{align*}
\Pr_{\sigma \mathcal{M}}(\Diamond E) & = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \\
\Pr_{\sigma \mathcal{M}}(\Diamond E \mid \Diamond c) & = \frac{1}{4}
\end{align*}
\]

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?

No

Consider the scheduler $\sigma$ that schedules $\alpha$ and $\beta$ with probability 1/2 in state $i$.

$$\Pr_{\mathcal{M}}^{\sigma}(\Diamond E) = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8}$$
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 $\sigma$ that schedules $\alpha$ and $\beta$ with probability $1/2$ in state $i$.

$$
\Pr_{\mathcal{M}}^\sigma(\diamond E) = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8} > \frac{1}{2} = \Pr_{\mathcal{M}}^\sigma(\diamond E | \diamond c)
$$
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 $\sigma$ that schedules $\alpha$ and $\beta$ with probability $1/2$ in state $i$.

$$\Pr_{\mathcal{M}}(\Diamond E) = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8} > \frac{1}{2} = \Pr_{\mathcal{M}}(\Diamond E | \Diamond c)$$
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 deterministic schedulers $\sigma_\alpha$ and $\sigma_\beta$ that schedule $\alpha$ resp. $\beta$ in state $i$. 

Consequence: Randomization needed for refuting the PR condition!
Example: PR cause in MDP

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

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

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

No, although PR condition holds for all deterministic schedulers

Consider the deterministic schedulers $\sigma_\alpha$ and $\sigma_\beta$ that schedule $\alpha$ resp. $\beta$ in state $i$.
e

$\sigma_\alpha$ 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 $\sigma_\alpha$ and $\sigma_\beta$ that schedule $\alpha$ resp. $\beta$ in state $i$.

$\sigma_\alpha$ irrelevant for PR condition as state $c$ is not reachable

$$\Pr^\sigma_{\mathcal{M}}(\Diamond E) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} < \frac{1}{2} = \Pr^\sigma_{\mathcal{M}}(\Diamond E | \Diamond c)$$
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 deterministic schedulers $\sigma_\alpha$ and $\sigma_\beta$ that schedule $\alpha$ resp. $\beta$ 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 $\sigma$ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1 - \lambda$. 
Example: PR cause in MDP

Let $\sigma$ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1 - \lambda$.

If $\lambda = 1$ then $\sigma$ is irrelevant (as $C$ is not reachable along $\sigma$-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 $\sigma$ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1 - \lambda$.

If $\lambda = 1$ then $\sigma$ is irrelevant (as $C$ is not reachable along $\sigma$-paths).
Otherwise:

$$
\Pr_{\mathcal{M}}^\sigma(\Diamond E) = \frac{1}{4} \cdot \lambda + \frac{1}{2} \cdot \frac{1}{2} \cdot (1 - \lambda) = \frac{1}{4}
$$
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 $\sigma$ be a scheduler with $\sigma(i)(\alpha) = \lambda$ and $\sigma(i)(\beta) = 1 - \lambda$.

If $\lambda = 1$ then $\sigma$ is irrelevant (as $c$ is not reachable along $\sigma$-paths).
Otherwise:

$$\Pr_{\mathcal{M}}^{\sigma}(\Diamond E) = \frac{1}{4} \cdot \lambda + \frac{1}{2} \cdot \frac{1}{2} \cdot (1 - \lambda) = \frac{1}{4} < \frac{1}{2} = \Pr_{\mathcal{M}}^{\sigma}(\Diamond E | \Diamond c)$$
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.
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.
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 $\text{CCC}$ is given: 4 types of terminal states
  - covered effect states: only accessible via $\text{CCC}$
  - uncovered effect states: not accessible from $\text{CCC}$
  - noneffect terminal states after $\text{CCC}$: only accessible via $\text{CCC}$
  - other noneffect terminal states: not accessible from $\text{CCC}$

And each $c \in \text{Cc} \in C$ has a single action with terminal successors (a covered effect state with prob. $p_c = \text{Pr}_{\text{min}}(\Box E)$) and a noneffect state with prob. $1 - p_c = \text{Pr}_{\text{min}}(\Box E)$.
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 $C$ 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 $c \in Cc \in Cc \in Cc \in C$ has a single action with terminal successors

$\Rightarrow$ covered effect states: only accessible via $C$ (TP)
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. $p_c = \Pr_c(\Diamond E)$ 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

Task: Given $\text{MMM}$, $\text{EEE}$, $\text{CCC}$, check whether $\text{CCC}$ is an SPR cause.

Observation: $\text{CCC}$ is an SPR cause iff for each state $c \in C$, $\{c\}$ is an SPR cause.
Checking the SPR condition

Task: Given $\mathcal{M}, E, C$, check whether $C$ 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 $p_c = \Pr_{\mathcal{M},c}^{\min}(\Diamond E)$. 
Task: Given $\mathcal{M}, E, c$, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP where $p_c = \Pr_{\mathcal{M},c}(\Diamond E)$.

$$p_c = \Pr_{\mathcal{N},c}(\Diamond E | \Diamond c)$$

for each scheduler $\sigma$ 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 $p_c = \Pr^\text{min}_{\mathcal{M}, c}(\Diamond E)$.

Let $q = \Pr^\text{max}_{\mathcal{N}}(\Diamond E)$.

$$p_c = \Pr^\sigma_{\mathcal{N}, c}(\Diamond E \mid \Diamond c)$$ for each scheduler $\sigma$ in $\mathcal{N}$ that reaches $c$.
Task: Given $\mathcal{M}, E, c$, check whether $\{c\}$ is an SPR cause.

Let $\mathcal{N}$ be the transformed MDP where $p_c = \Pr_{\mathcal{M}, c}^{\min}(\Diamond E)$.

Let $q = \Pr_{\mathcal{N}}^{\max}(\Diamond E)$.

If $q < p_c$: SPR condition holds.

If $q > p_c$: SPR condition does not hold.

$$p_c = \Pr_{\mathcal{N}, c}^{\sigma}(\Diamond E | \Diamond c)$$

for each scheduler $\sigma$ in $\mathcal{N}$ that reaches $c$. 
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 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 $p_c = \Pr_{\mathcal{M}, c}^\text{min}(\Diamond E)$. Let $q = \Pr_{\mathcal{N}}^\text{max}(\Diamond E)$.

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:

\[ C \text{ violates the GPR condition} \iff \text{there is an MR-scheduler}\]

\[ \text{refuting the GPR condition} \]

• linear balance equations for the expected frequencies:

\[ \Pr(\Box x s t \equiv t_0) = \Pr(x s t \mid \neg E) \]

\[ \forall s t \in C : \quad \Pr(x s t) \geq FN \]

and

\[ \Pr(\Box s t \mid \neg E) = \Pr(x s t) - FN \]
Checking the GPR condition

After the model transformation:

\[ C \text{ violates the GPR condition } \iff \begin{cases} \text{there is an MR-scheduler} \\ \text{refuting the GPR condition} \end{cases} \]

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 \((s, \alpha)\)

under such an MR-scheduler violating the GPR condition
Checking the GPR condition

- linear balance equations for the expected frequencies:

\[ x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, t) \quad \text{for each non-initial state } t \]

\[ x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, s_0) + 1 \quad \text{for the initial state } s_0 \]
Checking the GPR condition

- linear balance equations for the expected frequencies:
  \[ x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, t) \]
  for each non-initial state \( t \)
  \[ x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, s_0) + 1 \]
  for the initial state \( s_0 \)

- quadratic constraint for the violation of the GPR-condition:
  \[ x_C \cdot x_{FN} \geq (1-x_C) \cdot \sum_{s \in C} x_s \cdot p_s \]

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)
Checking the GPR condition

- linear balance equations for the expected frequencies:
  \[ x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, t) \quad \text{for each non-initial state } t \]
  \[ x_{s0} = \sum_{\alpha} x_{s0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, s_0) + 1 \quad \text{for the initial state } s_0 \]

- quadratic constraint for the violation of the GPR-condition:
  \[ x_C \cdot x_{FN} \geq (1 - x_C) \cdot \sum_{s \in C} x_s \cdot p_s \]

  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)

\[
Pr(\Diamond E | \Diamond C) = \frac{\sum_{s \in C} x_s \cdot p_s}{x_C}
\]
Checking the GPR condition

- linear balance equations for the expected frequencies:
  
  \[ x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, t) \quad \text{for each non-initial state } t \]
  
  \[ x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, s_0) + 1 \quad \text{for the initial state } s_0 \]

- quadratic constraint for the violation of the GPR-condition:
  
  \[ x_C \cdot x_{FN} \geq (1-x_C) \cdot \sum_{s \in C} x_s \cdot p_s \]

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)

\[
\Pr(\Diamond E | \Diamond C) = \frac{\sum_{s \in C} x_s \cdot p_s}{x_C} \quad \text{and} \quad \Pr(\Diamond E | \neg \Diamond C) = \frac{x_{FN}}{1-x_C}
\]
Checking the GPR condition

- linear balance equations for the expected frequencies:
  \[
  x_t = \sum_{\alpha} x_{t,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, t) \quad \text{for each non-initial state } t
  \]
  \[
  x_{s_0} = \sum_{\alpha} x_{s_0,\alpha} = \sum_{s,\alpha} x_{s,\alpha} \cdot P(s, \alpha, s_0) + 1 \quad \text{for the initial state } s_0
  \]

- quadratic constraint for the violation of the GPR-condition:
  \[
  x_C \cdot x_{FN} \geq (1-x_C) \cdot \sum_{s \in C} x_s \cdot p_s
  \]
  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 \( x_s,\alpha \geq 0 \) 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

- for fixed effect set
- take inspiration of quality measures used in statistical analysis for a good coverage 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

- for fixed effect set $E$ and GPR cause $C$
- take inspiration of quality measures used in statistical analysis for a good coverage of effect scenarios
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 coverage 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”)

\[ \text{prec}(C) = \inf_{\sigma} \Pr_{\mathcal{M}}^{\sigma}(\lozenge E \mid \lozenge C) \]

ranges over all schedulers
with \( \Pr_{\mathcal{M}}^{\sigma}(\lozenge C) > 0 \)

\[ \frac{TP}{TP + FP} \]
Quality measures for causes

precision (accuracy for “(true or false) positives”)

\[ \text{prec}(C) = \inf_{\sigma} \Pr_{\mathcal{M}}( \diamond E \mid \diamond C ) \]

recall (sensitivity):

\[ \text{recall}(C) = \inf_{\sigma} \Pr_{\mathcal{M}}( \diamond C \mid \diamond E ) \]

ranges over all schedulers
with \( \Pr_{\mathcal{M}}(\diamond E) > 0 \)
Quality measures for causes

precision (accuracy for “(true or false) positives”)

\[ \text{prec}(C) = \inf \Pr_{\mathcal{M}}^{\sigma}(\Diamond E \mid \Diamond C) \]

recall (sensitivity):

\[ \text{recall}(C) = \inf \Pr_{\mathcal{M}}^{\sigma}(\Diamond C \mid \Diamond E) \]

covratio (fraction of covered and uncovered effects)

\[ \text{covrat}(C) = \inf \frac{\Pr_{\mathcal{M}}^{\sigma}(\Diamond C \land \Diamond E)}{\Pr_{\mathcal{M}}^{\sigma}((\neg C) \cup E)} \]

ranges over all schedulers with \( \Pr_{\mathcal{M}}^{\sigma}((\neg C) \cup E) > 0 \)

\[ \frac{\text{TP}}{\text{TP} + \text{FP}} \]

\[ \frac{\text{TP}}{\text{TP} + \text{FN}} \]

\[ \frac{\text{TP}}{\text{FN}} \]
Quality measures for causes

precision (accuracy for “(true or false) positives”)
\[
prec(C) = \inf_\sigma \Pr_M^\sigma(\Diamond E | \Diamond C )
\]

recall (sensitivity):
\[
\text{recall}(C) = \inf_\sigma \Pr_M^\sigma(\Diamond C | \Diamond E )
\]

coverage ratio (fraction of covered and uncovered effects)
\[
covrat(C) = \inf_\sigma \frac{\Pr_M^\sigma(\Diamond C \land \Diamond E)}{\Pr_M^\sigma((\neg C) \cup E)}
\]

f-score (harmonic mean of precision and recall)
\[
fscore(C) = \inf_\sigma \frac{prec^\sigma(C) \cdot recall^\sigma(C)}{prec^\sigma(C) + recall^\sigma(C)}
\]
Quality measures for causes

precision (accuracy for "(true or false) positives")

\[ \text{prec}(C) = \inf_{\sigma} \Pr_{M}^{\sigma}(\lozenge E | \lozenge C) \]

recall (sensitivity):

\[ \text{recall}(C) = \inf_{\sigma} \Pr_{M}^{\sigma}(\lozenge C | \lozenge E) \]

coverage ratio (fraction of covered and uncovered effects)

\[ \text{covrat}(C) = \inf_{\sigma} \frac{\Pr_{M}^{\sigma}(\lozenge C \land \lozenge E)}{\Pr_{M}^{\sigma}((\neg C) \cup E)} \]

f-score (harmonic mean of precision and recall)

\[ \text{fscore}(C) = \inf_{\sigma} \frac{\text{prec}^{\sigma}(C) \cdot \text{recall}^{\sigma}(C)}{\text{prec}^{\sigma}(C) + \text{recall}^{\sigma}(C)} \]

already taken into account in the GPR condition; precision says nothing about coverage

\[ \frac{\text{TP}}{\text{TP} + \text{FN}} \]

computing precision & recall: via standard techniques for condition prob. in MDPs

computing covrat & f-score: via reduction to SSPP (stoch. shortest path problem)
Quality measures for causes

precision (accuracy for "(true or false) positives")

\[ prec(C) = \inf_{\sigma} \Pr^\sigma_M(\Diamond E \mid \Diamond C) \]

recall (sensitivity):

\[ recall(C) = \inf_{\sigma} \Pr^\sigma_M(\Diamond C \mid \Diamond E) \]

coverage ratio (fraction of covered and uncovered effects)

\[ \text{covrat}(C) = \inf_{\sigma} \frac{\Pr^\sigma_M(\Diamond C \land \Diamond E)}{\Pr^\sigma_M((\neg C) \cup E)} \]

f-score (harmonic mean of precision and recall)

\[ \text{fscore}(C) = \inf_{\sigma} \frac{\text{prec}^\sigma(C) \cdot \text{recall}^\sigma(C)}{\text{prec}^\sigma(C) + \text{recall}^\sigma(C)} \]

computing precision & recall:

via standard techniques for condition prob. in MDPs

computing covrat & f-score:

via reduction to SSPP (stoch. shortest path problem)
Quality measures for causes

precision (accuracy for “(true or false) positives”)

\[ \text{prec}(C) = \inf_{\sigma} \Pr_{\mathcal{M}}^\sigma(\Diamond E \mid \Diamond C) \]

recall (sensitivity):

\[ \text{recall}(C) = \inf_{\sigma} \Pr_{\mathcal{M}}^\sigma(\Diamond C \mid \Diamond E) \]

coverage ratio (fraction of covered and uncovered effects)

\[ \text{covrat}(C) = \inf_{\sigma} \frac{\Pr_{\mathcal{M}}^\sigma(\Diamond C \land \Diamond E)}{\Pr_{\mathcal{M}}^\sigma((\neg C) \cup E)} \]

f-score (harmonic mean of precision and recall)

\[ \text{fscore}(C) = \inf_{\sigma} \frac{\text{prec}^\sigma(C) \cdot \text{recall}^\sigma(C)}{\text{prec}^\sigma(C) + \text{recall}^\sigma(C)} \]

computing precision & recall:
via standard techniques for condition prob. in MDPs

computing covrat & f-score:
via reduction to SSPP (stoch. shortest path problem)
Coverage ratio and f-score

Coverage ratio (fraction of covered and uncovered effects)

covrat(C) = \infty

σPrM(♢C ∧ ♢E)

σPrM((¬C) U E) \infty

σPrM((¬C) U E) = \infty

σTPσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσσsigma
Coverage ratio and f-score

coverage ratio (fraction of covered and uncovered effects)

\[ \text{covrat}(C) = \inf_{\sigma} \frac{\Pr^\sigma_M(\Diamond C \land \Diamond E)}{\Pr^\sigma_M((\neg C) \cup E)} = \inf_{\sigma} \frac{\text{TP}^\sigma}{\text{FN}^\sigma} \]

f-score (harmonic mean of precision and recall)

\[ \text{fscore}(C) = \inf_{\sigma} \frac{\text{prec}^\sigma(C) \cdot \text{recall}^\sigma(C)}{\text{prec}^\sigma(C) + \text{recall}^\sigma(C)} \]

TP  true positive (covered effects)
FN  false negative (uncovered effects)
Coverage ratio and f-score

Coverage ratio (fraction of covered and uncovered effects)

$$covrat(C) = \inf_{\sigma} \frac{Pr^\sigma_M(\Diamond C \land \Diamond E)}{Pr^\sigma_M((\neg C) \cup E)} = \inf_{\sigma} \frac{TP^\sigma}{FN^\sigma}$$

F-score (harmonic mean of precision and recall)

$$fscore(C) = \frac{2}{X + 2} \quad \text{where} \quad X = \sup_{\sigma} \frac{FP^\sigma + FN^\sigma}{TP^\sigma}$$

TP true positive (covered effects)  TN true negative (noeffect without C)
FN false negative (uncovered effects)  FP false positive (noeffect after C)
Coverage ratio and f-score

Coverage ratio (fraction of covered and uncovered effects)

\[
\text{covrat}(C) = \inf_{\sigma} \frac{\Pr_{M}^{\sigma}(\Box C \land \Box E)}{\Pr_{M}^{\sigma}((\neg C) \cup E)} = \inf_{\sigma} \frac{\text{TP}^{\sigma}}{\text{FN}^{\sigma}}
\]

f-score (harmonic mean of precision and recall)

\[
\text{fscore}(C) = \frac{2}{\frac{\text{X}}{2} + 2} \quad \text{where} \quad \text{X} = \sup_{\sigma} \frac{\text{FP}^{\sigma} + \text{FN}^{\sigma}}{\text{TP}^{\sigma}}
\]

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)

\[ \text{covrat}(C) = \inf_{\sigma} \frac{\Pr_M(\Diamond C \land \Diamond E)}{\Pr_M((\neg C) \cup E)} = \inf_{\sigma} \frac{\text{TP}^\sigma}{\text{FN}^\sigma} \]

F-score (harmonic mean of precision and recall)

\[ \text{fscore}(C) = \frac{2}{X + 2} \text{ where } X = \sup_{\sigma} \frac{\text{FP}^\sigma + \text{FN}^\sigma}{\text{TP}^\sigma} \]

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 \( \frac{\Pr_M(\Diamond U)}{\Pr_M(\Diamond V)} \) quotient of probabilities for reaching disjoint sets of terminal states
After model transformation ...

![Diagram showing model transformation with symbols for true negatives (TN), false positives (FP), true positives (TP), and coverage ratio (Cov)].

Coverage ratio (Cov) = \frac{TP}{TP + FP}

Coverage fraction (Frac) = \frac{TP}{TP + FP + FN}

Support (Sup) = \frac{TP}{TP + FP}

Recall (Rec) = \frac{TP}{TP + FN}

Precision (Pre) = \frac{TP}{TP + FP}

F-score (Fscore) = \frac{2 \times Pre \times Rec}{Pre + Rec}
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

\[
\text{covrat}(C) = \inf_{\sigma} \frac{TP^\sigma}{FN^\sigma}
\]
After model transformation ...

coverage ratio
fraction of covered and uncovered effects

\[ \text{covrat}(\mathcal{C}) = \inf_{\sigma} \frac{TP_\sigma}{FN_\sigma} \]

f-score
harmonic mean of precision & recall

\[ \text{fscore}(\mathcal{C}) = \frac{2}{X + 2} \]

where \( X = \sup_{\sigma} \frac{FP_\sigma + FN_\sigma}{TP_\sigma} \)
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 $\inf_{\sigma} \frac{\Pr_{\mathcal{M}}(\Diamond U)}{\Pr_{\mathcal{M}}(\Diamond V)}$

(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 $\inf_{\sigma} \frac{\Pr_{M}^{\sigma}(\Diamond U)}{\Pr_{M}^{\sigma}(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP
weight 1 for $U$, weight 0 for all other states
Covratio and f-score via SSPP

Given MDP $\mathcal{M}$
- without end components
- $U, V$ disjoint sets of terminal states

Goal: compute $\inf_{\sigma} \frac{\Pr_{\mathcal{M}}(\Diamond U)}{\Pr_{\mathcal{M}}(\Diamond V)}$

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.
Given MDP $\mathcal{M}$
- without end components
- $U$, $V$ disjoint sets of terminal states

Goal: compute $\inf_{\sigma} \frac{\Pr_{\mathcal{M}}(\Diamond U)}{\Pr_{\mathcal{M}}(\Diamond V)}$

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.
Given MDP $\mathcal{M}$
- without end components
- $U, V$ disjoint sets of terminal states

Goal: compute $\inf_{\sigma} \frac{\Pr^\sigma_M(\Diamond U)}{\Pr^\sigma_M(\Diamond V)}$

Let $\mathcal{N}$ be the transformed weighted MDP
weight 1 for $U$, weight 0 for all other states

$$\inf_{\sigma} \frac{\Pr^\sigma_M(\Diamond U)}{\Pr^\sigma_M(\Diamond V)} = \inf_{\sigma} \mathbb{E}^\sigma_{\mathcal{N}}(\text{“accumulated weight until reaching } V\text{”})$$

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
- Optimization 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):

- 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

"canonical SPR cause": \(\text{CCC} = \text{union of all singleton SPR causes}\)

- recall-optimal: obvious as any SPR is a subset of \(\text{CCC}\)
- covratio-opt = recall-opt: \(TP < FN \iff TP'FN + TP < FN' + TP'\)
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)
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: $\frac{TP}{FN} < \frac{TP'}{FN'}$ iff $\frac{TP}{FN+TP} < \frac{TP'}{FN'+TP'}$
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 $\mathcal{M}$

set of states $c, d, ..., C = \{c, d, ..., \}$

$p_c = \Pr(c(\diamond E)) > \Pr_M(\diamond E)$

nondeterministic choice in CCC

action $\alpha$: "$c$ selected for SPR cause"

transition with prob. $p_c$
to new effect state

action $\beta$: "$c$ not selected for SPR cause"

reset transitions from TP, FN, FP

weight 1 for FN and FP

weight 0 for all other states
F-score optimal SPR cause in MC

\[ C = \{ c, d, \ldots \} \]

set of states \( c \) with
\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

MC \( \mathcal{M} \)

nondeterministic choice in \( \mathcal{C} \)-states
action \( \alpha \): "\( c \) selected for SPR cause"
move with prob. \( p_c \) to new effect state
with prob. \( 1 - p_c \) to a terminal non-effect state

action \( \beta \): "\( c \) not selected for SPR cause"
reset transitions from TP, FN, FP
weight 1 for FN and FP
weight 0 for all other states
F-score optimal SPR cause in MC

MDP $\mathcal{N}$

- $C = \{c, d, \ldots\}$
- set of states $c$ with $p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E)$

nondeterministic choice in $C$-states

- action $\alpha$: “$c$ selected for SPR cause”
- move with prob. $p_c$ to new effect state $\mathit{eff}$
- with prob. $1 - p_c$ to a terminal non-effect state
F-score optimal SPR cause in MC

MDP $\mathcal{N}$

\[ C = \{ c, d, \ldots \} \]

set of states $c$ with

\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

nondeterministic choice in $C$-states

action $\alpha$: “$c$ selected for SPR cause”
move with prob. $p_c$ to new effect state $\text{eff}$
with prob. $1-p_c$ to a terminal non-effect state

action $\beta$: “$c$ not selected for SPR cause”

reset transitions from TP, FN, FP
weight 1 for FN and FP
weight 0 for all other states
F-score optimal SPR cause in MC

\[ C = \{ c, d, \ldots \} \]

set of states \( c \) with
\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

nondeterministic choice in \( C \)-states

action \( \alpha \): “\( c \) selected for SPR cause”
move with prob. \( p_c \) to new effect state \( \text{eff} \) with prob. \( 1 - p_c \) to a terminal non-effect state

action \( \beta \): “\( c \) not selected for SPR cause”

\[ f_{\text{score}}(C) = \frac{2}{X_c + 2} \quad \text{where} \quad X_c = \frac{\text{FN}_c + \text{FP}_c}{\text{TP}_c} \]
F-score optimal SPR cause in MC

\[ C = \{c, d, \ldots\} \]

set of states \( c \) with
\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

nondeterministic choice in \( C \)-states
action \( \alpha \): “\( c \) selected for SPR cause”
move with prob. \( p_c \) to new effect state \( \text{eff} \)
with prob. \( 1 - p_c \) to a terminal non-effect state
action \( \beta \): “\( c \) not selected for SPR cause”
reset transitions from TP, FN, FP

\[ \text{fscore}(C) = \frac{2}{X_c + 2} \quad \text{where} \quad X_c = \frac{F_{Nc} + F_{Pc}}{T_{Pc}} \]
F-score optimal SPR cause in MC

\[ C = \{c, d, \ldots\} \]

set of states \( c \) with

\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

nondeterministic choice in \( C \)-states

action \( \alpha \): “\( c \) selected for SPR cause”

move with prob. \( p_c \) to new effect state \( \text{eff} \)

with prob. \( 1-p_c \) to a terminal non-effect state

action \( \beta \): “\( c \) not selected for SPR cause”

reset transitions from TP, FN, FP

weight 1 for FN and FP

weight 0 for all other states

\[ \text{fscore}(C) = \frac{2}{X_C+2} \]

where \( X_C = \frac{\text{FN}_C+\text{FP}_C}{\text{TP}_C} \)

\[ C = \{c, d, \ldots\} \]

set of states \( c \) with

\[ p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E) \]

nondeterministic choice in \( C \)-states

action \( \alpha \): “\( c \) selected for SPR cause”

move with prob. \( p_c \) to new effect state \( \text{eff} \)

with prob. \( 1-p_c \) to a terminal non-effect state

action \( \beta \): “\( c \) not selected for SPR cause”

reset transitions from TP, FN, FP

weight 1 for FN and FP

weight 0 for all other states

\[ \text{fscore}(C) = \frac{2}{X_C+2} \]

where \( X_C = \frac{\text{FN}_C+\text{FP}_C}{\text{TP}_C} \)
F-score optimal SPR cause in MC

$M = \{c, d, \ldots\}$

set of states $c$ with

$p_c = \Pr_c(\Diamond E) > \Pr_M(\Diamond E)$

nondeterministic choice in $C$-states

action $\alpha$: “$c$ selected for SPR cause”

move with prob. $p_c$ to new effect state $\text{eff}$

with prob. $1 - p_c$ to a terminal non-effect state

action $\beta$: “$c$ not selected for SPR cause”

reset transitions from TP, FN, FP

weight 1 for FN and FP

weight 0 for all other states

$$\max_C \text{fscore}(C) = \frac{2}{X+2} \quad \text{where} \quad X = E_{\mathcal{N}}^{\min}(\text{weight})$$
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 $C$:

<table>
<thead>
<tr>
<th></th>
<th>checking PR condition</th>
<th>computing quality measures (recall, coverage ratio, f-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR</td>
<td>$\in \mathbb{P}$</td>
<td>poly-time</td>
</tr>
</tbody>
</table>
| GPR   | MDP: $\in \text{PSPACE}$  
      | MC: $\in \mathbb{P}$                                       | 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:

<table>
<thead>
<tr>
<th></th>
<th>covratio-optimal = recall-optimal</th>
<th>f-score-optimal</th>
<th>threshold problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR</td>
<td>poly-time</td>
<td>MDP: poly-space</td>
<td>f-score threshold problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MC: poly-time</td>
<td>MDP: $\in$ NP $\cap$ coNP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MC: $\in$ P</td>
</tr>
<tr>
<td>GPR</td>
<td>poly-space</td>
<td></td>
<td>MDP: $\in$ PSPACE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MC: NP-complete</td>
</tr>
</tbody>
</table>
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
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]

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
  - 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 $\Pr(\text{effect} | \text{cause}) > \Pr(\text{effect} | \neg \text{cause})$

- many open questions: path events for causes and effects, other quality measures, backward causality, actionability, ...
Thank You