Actual Causality in Reactive Systems

Hadar Frenkel

February 2023

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

erc

Joint work with

Raimund Dachselt[2] Norine Coenen[1] Bernd Finkbeiner[1] Christopher Hahn[3]

Tom Horak[2] Niklas Metzger[1] Julian Siber[1]

[1]Cispa Helmholtz Center For Information Security
[2]Interactive Media Lab, TU Dresden
[3]Stanford University

Explaining Hyperproperty Violations

Norine Coenen[1](✉), Raimund Dachselt[2], Bernd Finkbeiner[1], Hadar Frenkel[1], Christopher Hahn[1], Tom Horak[3], Niklas Metzger[1], and Julian Siber[1]

[1] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
{norine.coenen,finkbeiner,hadar.frenkel,christopher.hahn, niklas.metzger,julian.siber}@cispa.de

[2] Interactive Media Lab, Technische Universität Dresden, Dresden, Germany
dachselt@acm.org

[3] elevait GmbH & Co. KG, Dresden, Germany
tom.horak@elevait.de

CAV Artifact Evaluation ★ Available

CAV Artifact Evaluation ★★ Functional

Abstract. Hyperproperties relate multiple computation traces to each other. Model checkers for hyperproperties thus return, in case a system model violates the specification, a set of traces as a counterexample. Fixing the erroneous relations between traces in the system that led to the counterexample is a difficult manual effort that highly benefits from additional explanations. In this paper, we present an explanation method for counterexamples to hyperproperties described in the specification logic HyperLTL. We extend Halpern and Pearl's definition of actual causality to sets of traces witnessing the violation of a HyperLTL formula, which allows us to identify the events that caused the violation. We report on the implementation of our method and show that it significantly improves on previous approaches for analyzing counterexamples returned by HyperLTL model checkers.

Temporal Causality in Reactive Systems

Norine Coenen[1], Bernd Finkbeiner[1], Hadar Frenkel[1], Christopher Hahn[2], Niklas Metzger[1], and Julian Siber[1](✉)

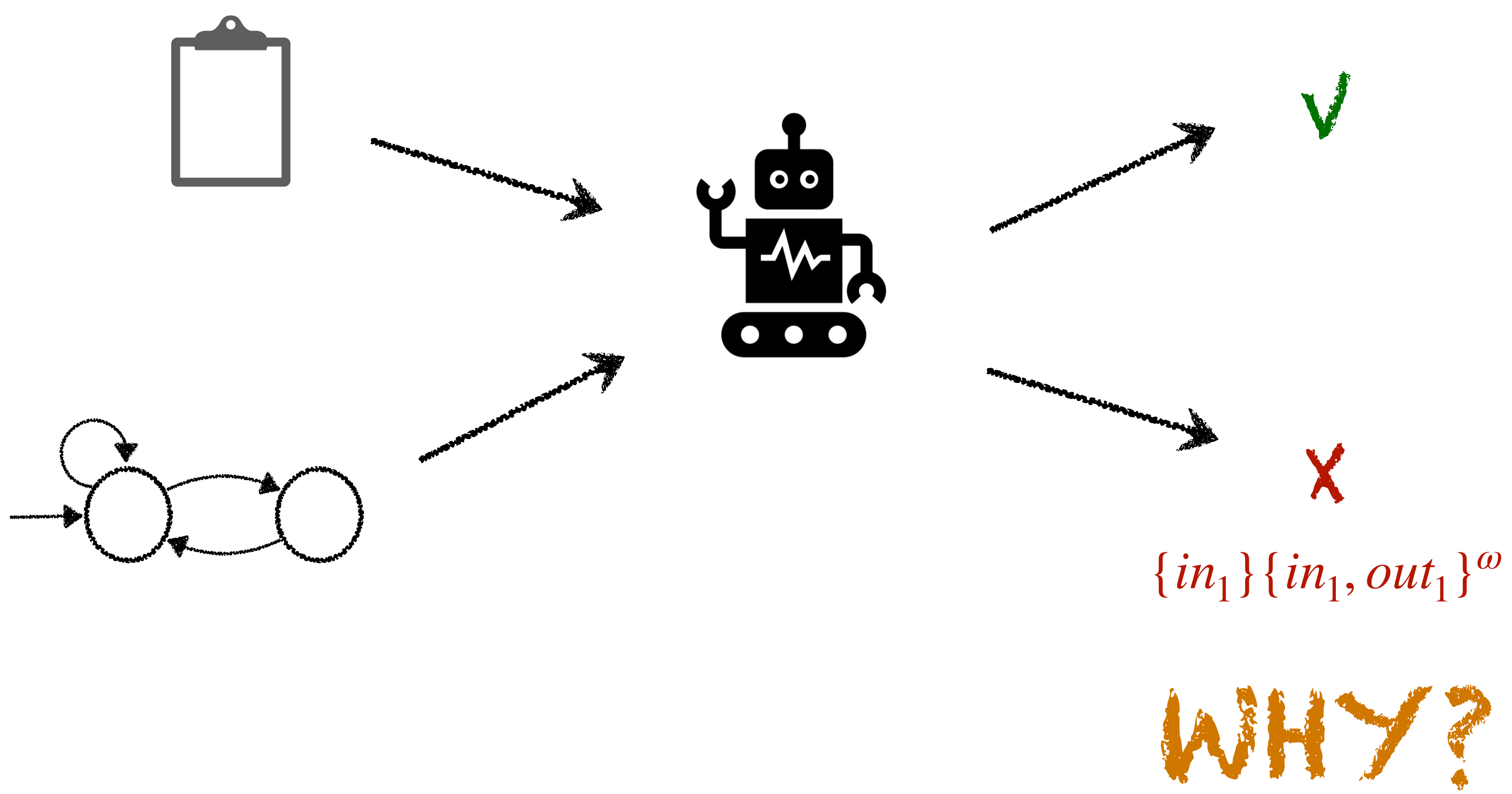
[1] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
{norine.coenen,finkbeiner,hadar.frenkel,niklas.metzger, julian.siber}@cispa.de

[2] Stanford University, Stanford, USA
hahn@cs.stanford.edu

Abstract. Counterfactual reasoning is an approach to infer what causes an observed effect by analyzing the hypothetical scenarios where a suspected cause is not present. The seminal works of Halpern and Pearl have provided a workable definition of counterfactual causality for finite settings. In this paper, we propose an approach to check causality that is tailored to reactive systems, i.e., systems that interact with their environment over a possibly infinite duration. We define causes and effects as trace properties which characterize the input and observed output behavior, respectively. We then instantiate our definitions for ω-regular properties and give automata-based constructions for our approach. Checking that an ω-regular property qualifies as a cause can then be encoded as a hyperproperty model-checking problem.

Explaining Hyperproperty Violations. Coenen, Dachselt, Finkbeiner, **F.**, Hahn, Horak, Metzger, Siber. (CAV 2022)
Temporal Causality in Reactive Systems. Coenen, Finkbeiner, **F.**, Hahn, Metzger, Siber. (ATVA 2022)

Model Checking

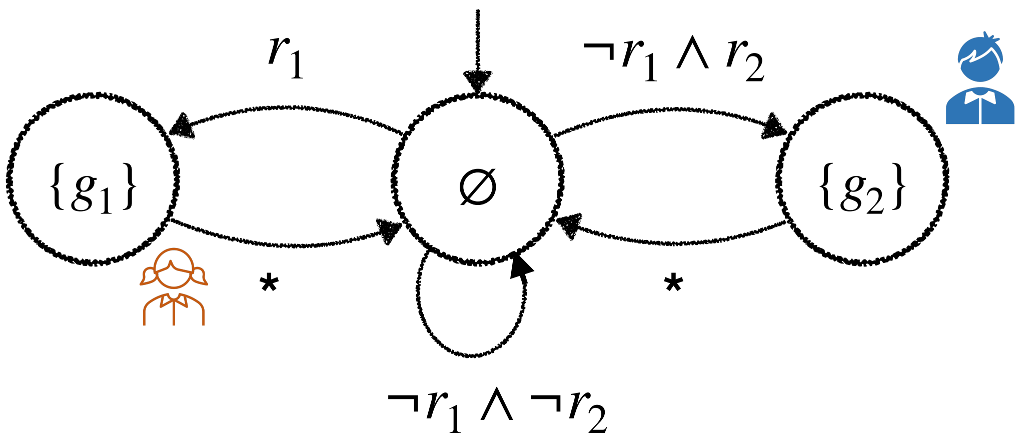


Model Checking

Reactive systems

Inputs trigger outputs

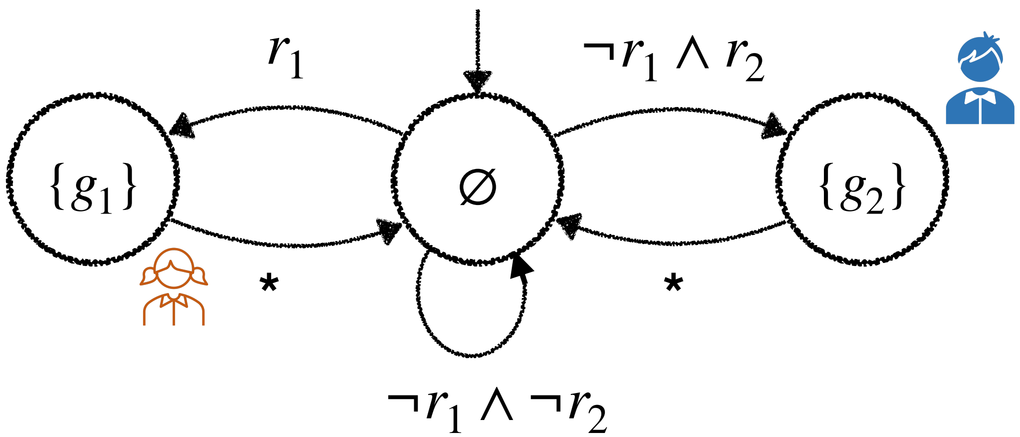
Infinite computations

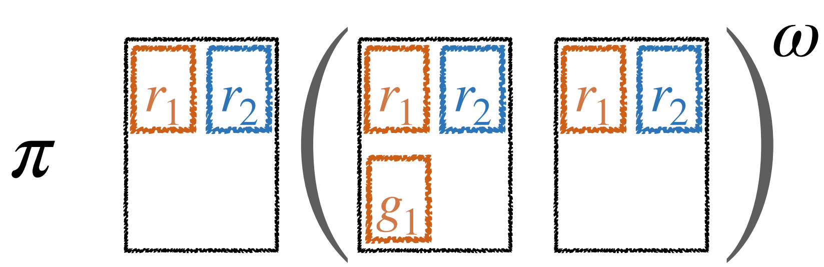


Model Checking

Causes over input sequences
Analyse the system dynamics

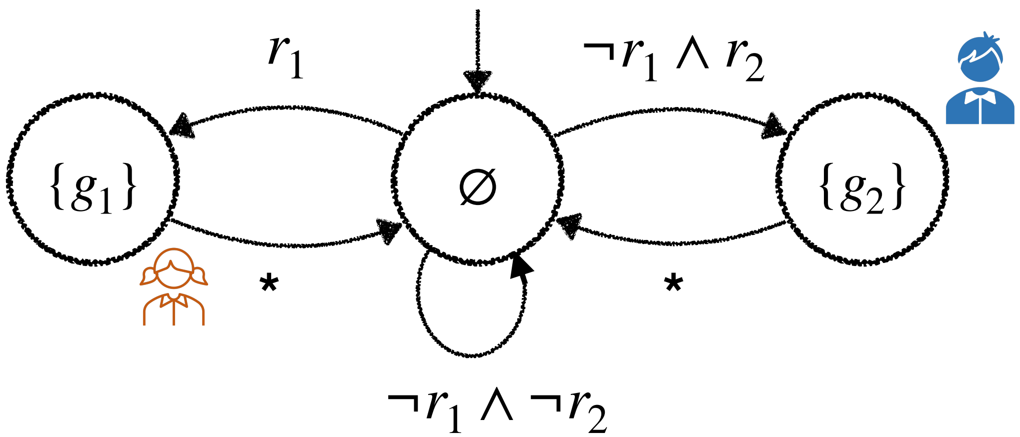
Is “always r_1” the cause for “always not g_2”?



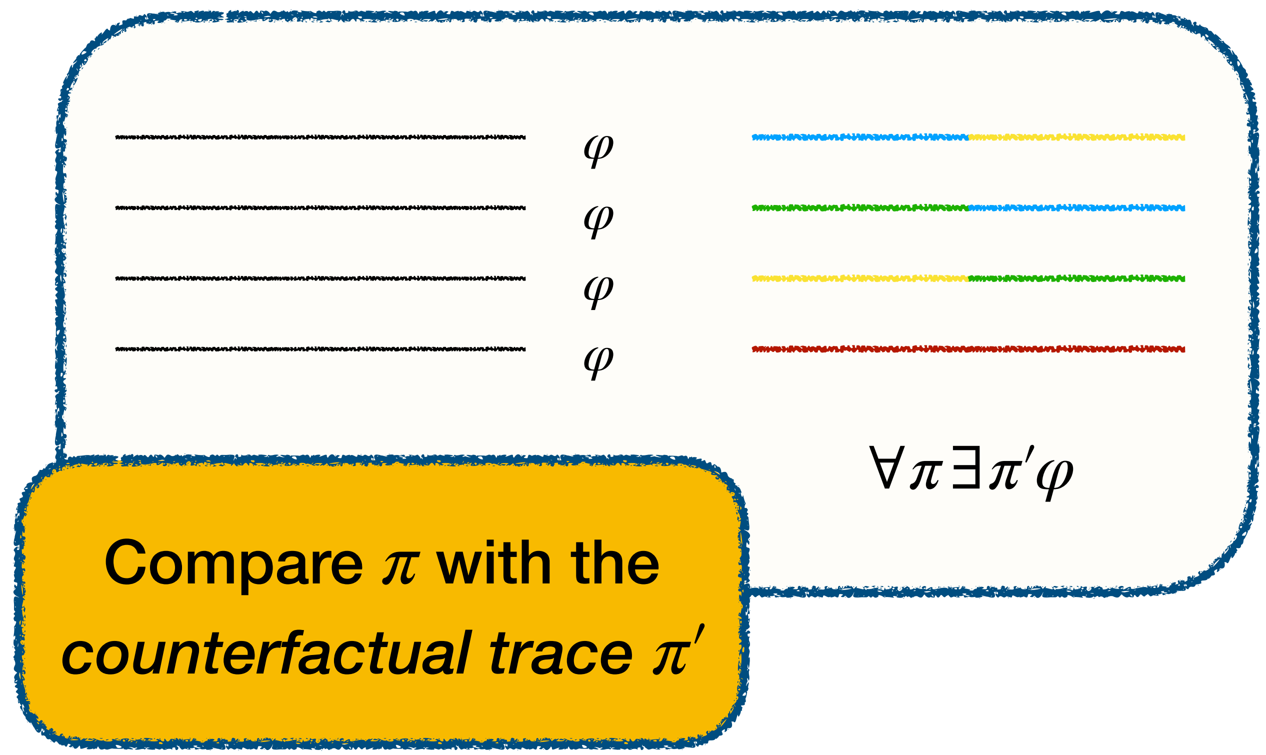


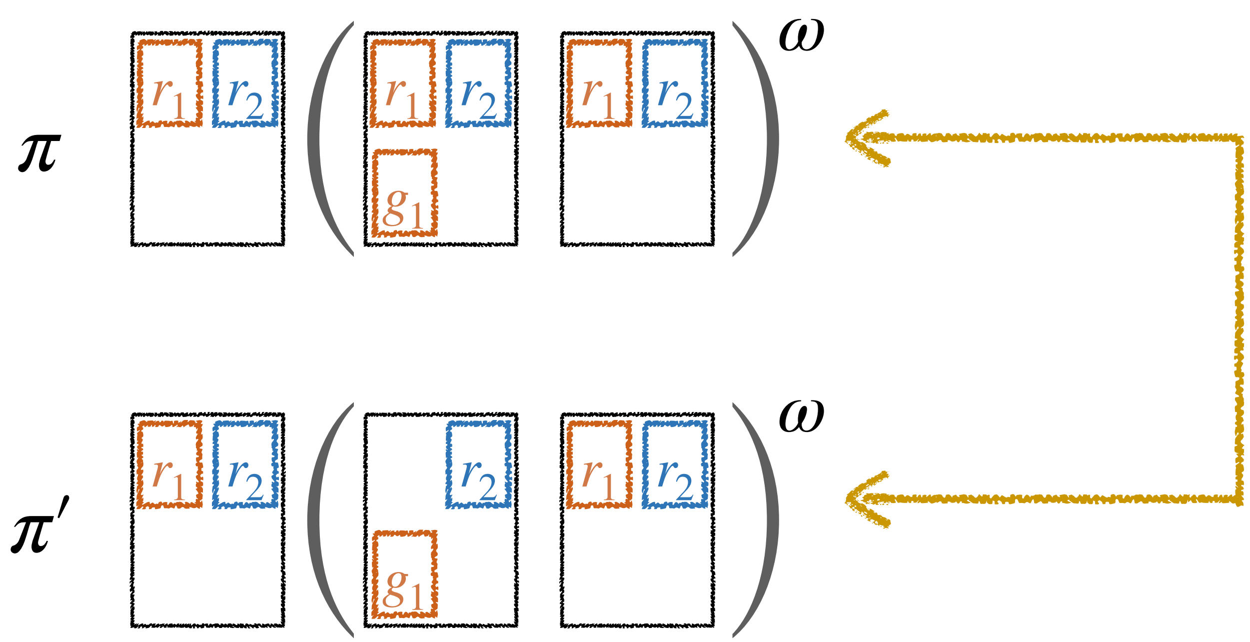
“eventually g_2”

Causality as a Hyperproperty



Is “always r_1”
the cause for “always not g_2”?





Actual Causality in Reactive Systems

1

$\{\langle p, \pi, 0\rangle\}$

Specific events on the trace that cause the violation

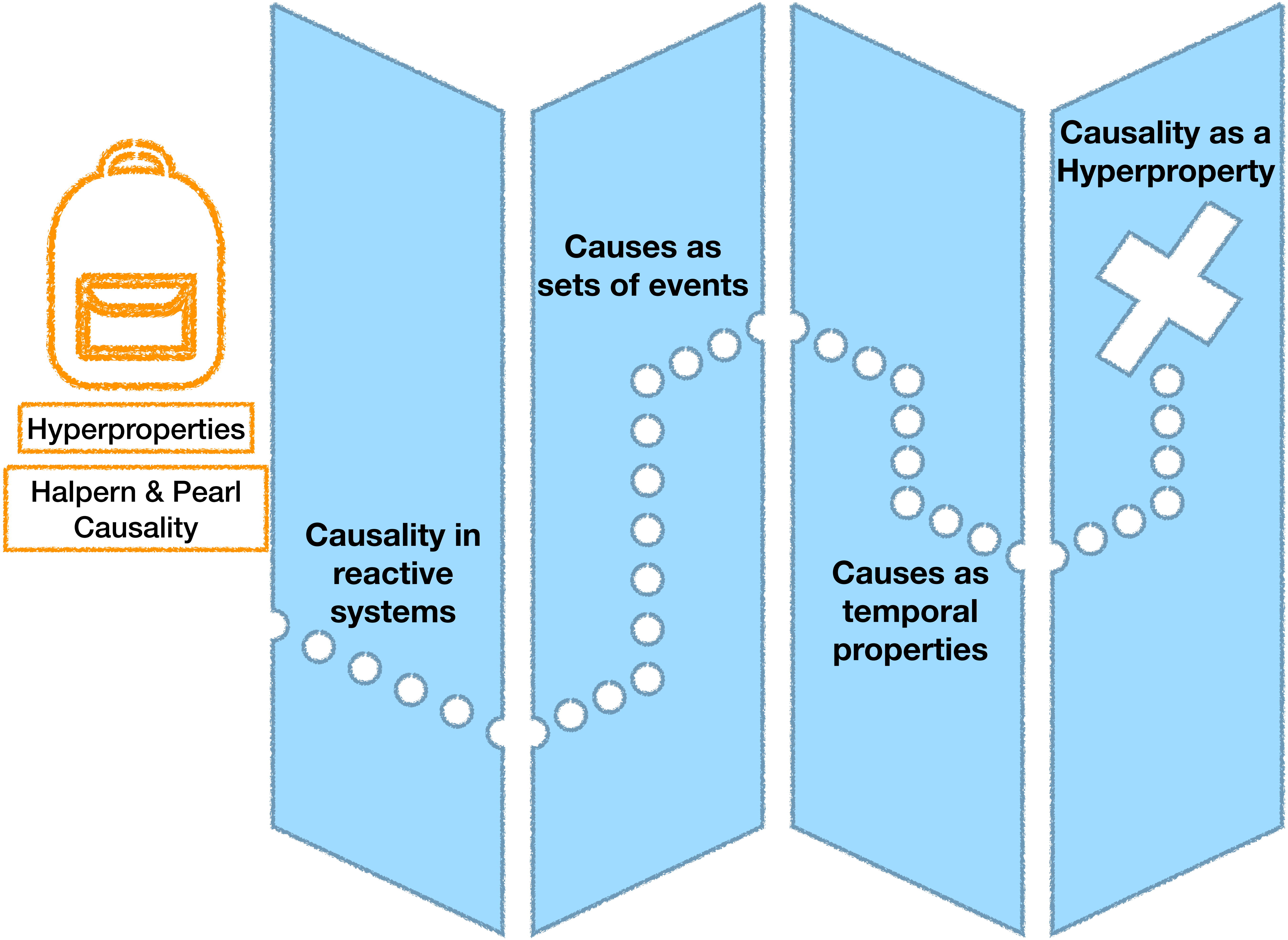
2

“infinitely often p”

Trace properties

Explainability — analysis of the counterexample

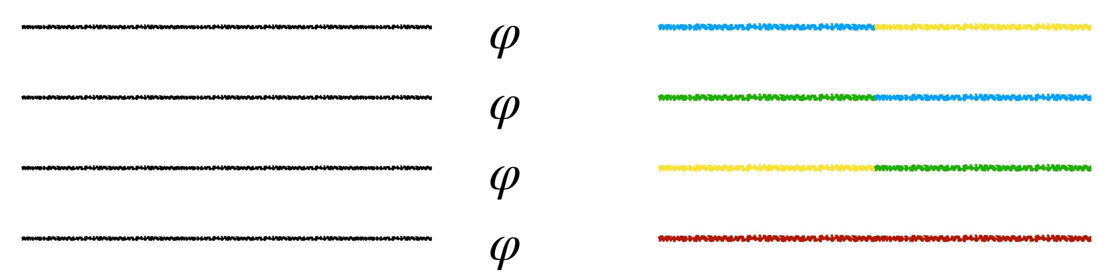
Applicability — repair



Hyperproperties
Halpern & Pearl Causality
Causality in reactive systems
Causes as sets of events
Causes as temporal properties
Causality as a Hyperproperty

Hyperproperties

- Extend trace properties (e.g., in LTL) to system properties
- Reason about sets of traces



$\forall \pi \exists \pi' \varphi$

Linear Temporal Logic — LTL

$\square p$ — p holds at every timepoint

$\lozenge p$ — p eventually holds

$\bigcirc p$ — p holds at the next timepoint

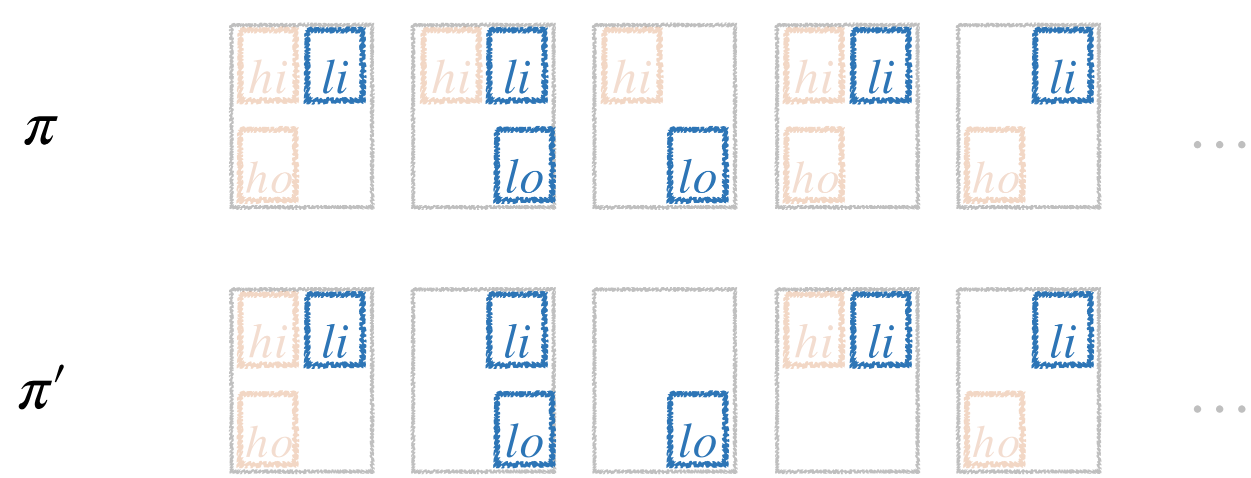
Hyperproperties. Clarkson and Schneider. (CSF 2008).

Hyperproperties

- Extend trace properties (e.g., in LTL) to system properties
- Reason about sets of traces
- Observational determinism:

HyperLTL — extending LTL with trace quantification

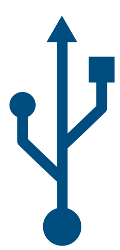
$$\forall \pi \forall \pi' \Box (li_\pi \leftrightarrow li_{\pi'}) \rightarrow \Box (lo_\pi \leftrightarrow lo_{\pi'})$$



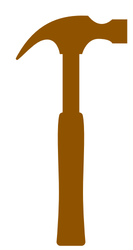
Hyperproperties. Clarkson and Schneider. (CSF 2008).
Temporal Logics for Hyperproperties. Clarkson, Finkbeiner, Koleini, Micinski, Rabe, and Sánchez. (POST 2014).

Hyperproperties

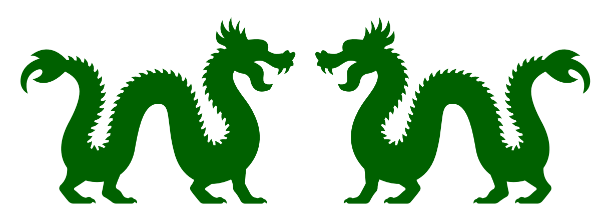
- Extend trace properties (e.g., in LTL) to system properties
- Reason about sets of traces



Information-flow properties [Observational determinism]



Robustness



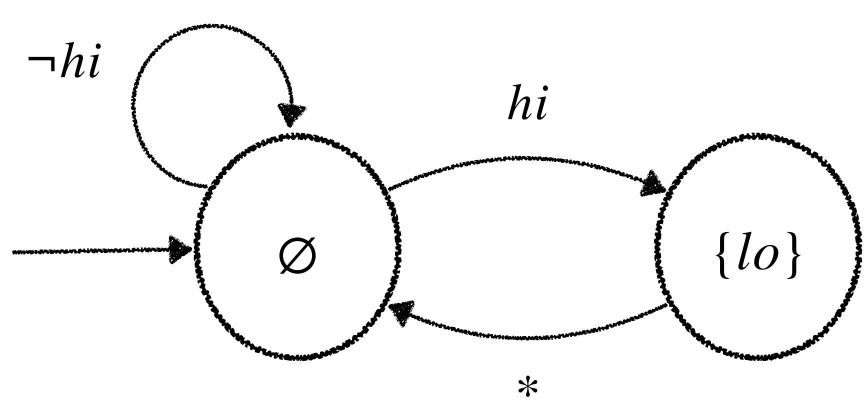
Symmetry

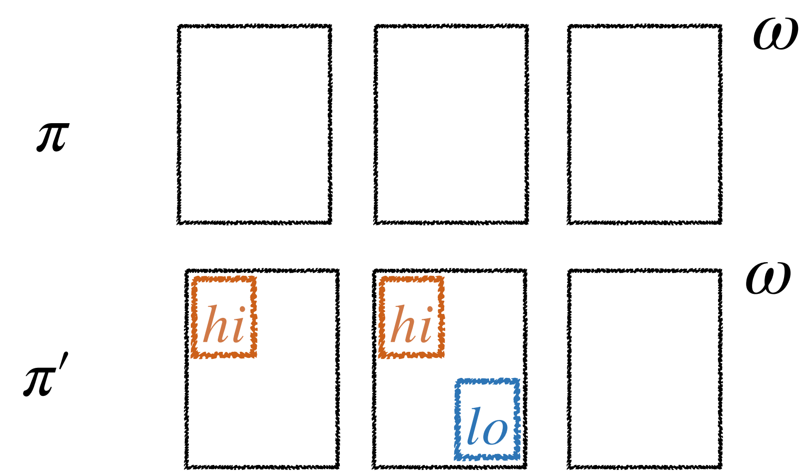
Causality

Hyperproperties. Clarkson and Schneider. (CSF 2008).

Explaining Hyperproperty Violations

$$\forall \pi \forall \pi' \Box (li_{\pi} \leftrightarrow li_{\pi'}) \rightarrow \Box (lo_{\pi} \leftrightarrow lo_{\pi'})$$





WHY?

Joseph Y. Halpern, Judea Pearl:
Causes and Explanations: A Structural-Model Approach: Part 1: Causes. UAI 2001: 194-202

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

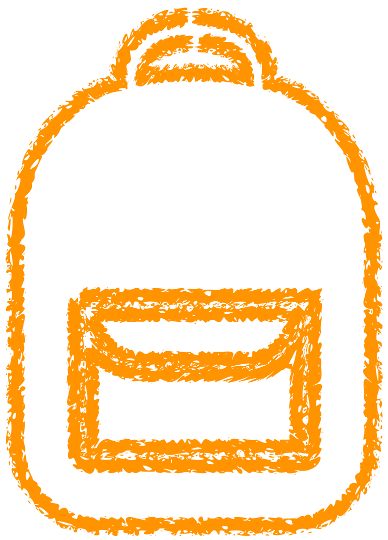
Author(s): Joseph Y. Halpern and Judea Pearl

Source: *The British Journal for the Philosophy of Science*, Dec., 2005, Vol. 56, No. 4 (Dec., 2005), pp. 843-887

Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

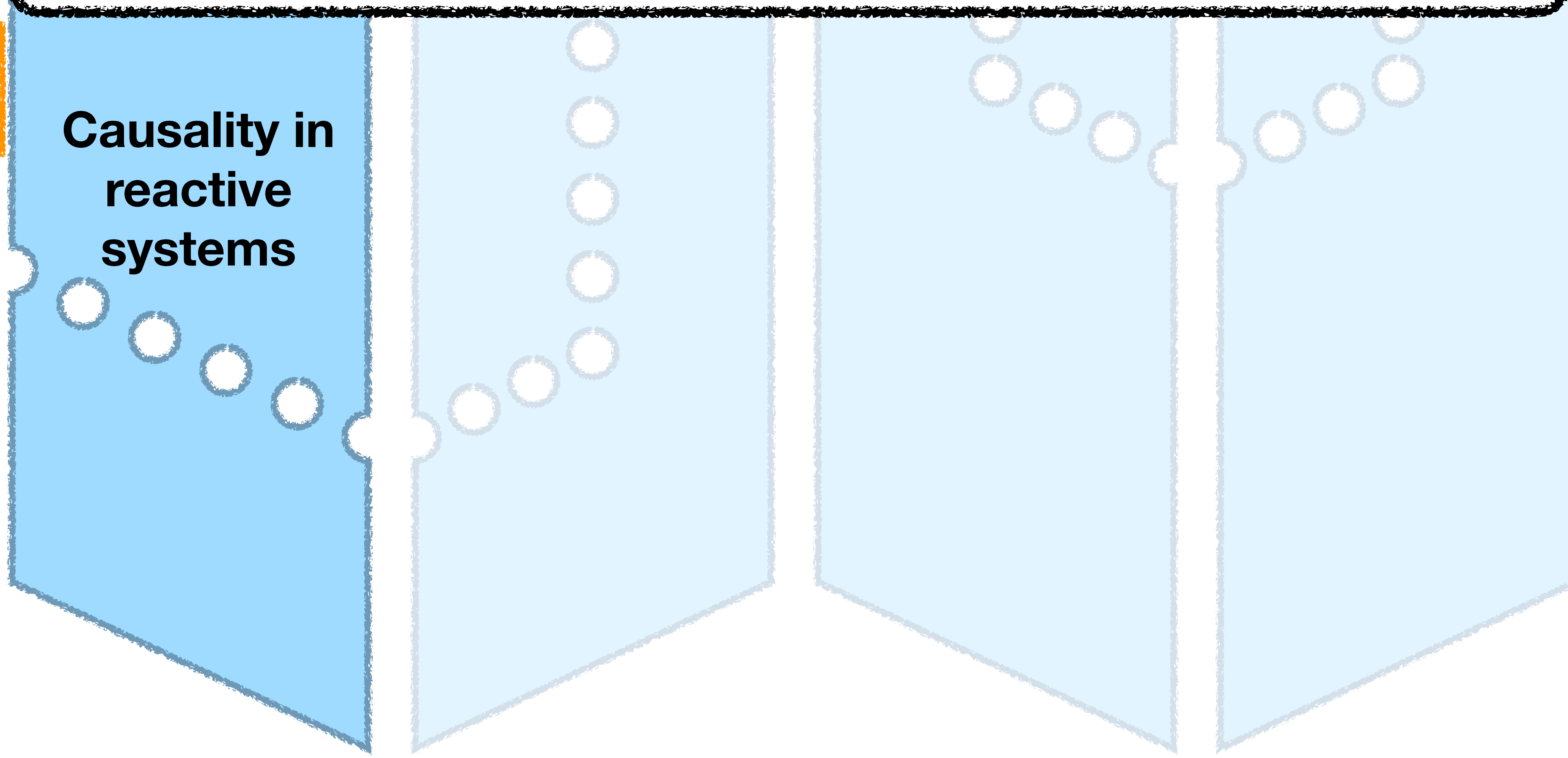
A Modification of the Halpern-Pearl Definition of Causality

Joseph Y. Halpern*
Cornell University

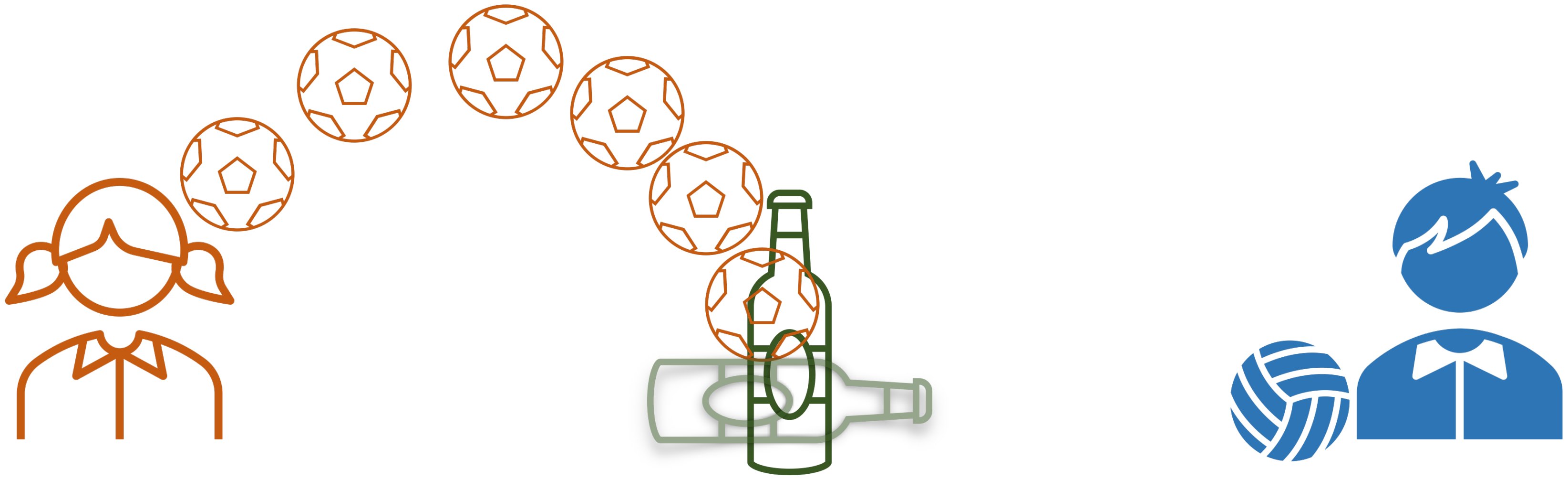


Hyperproperties

Halpern & Pearl Causality

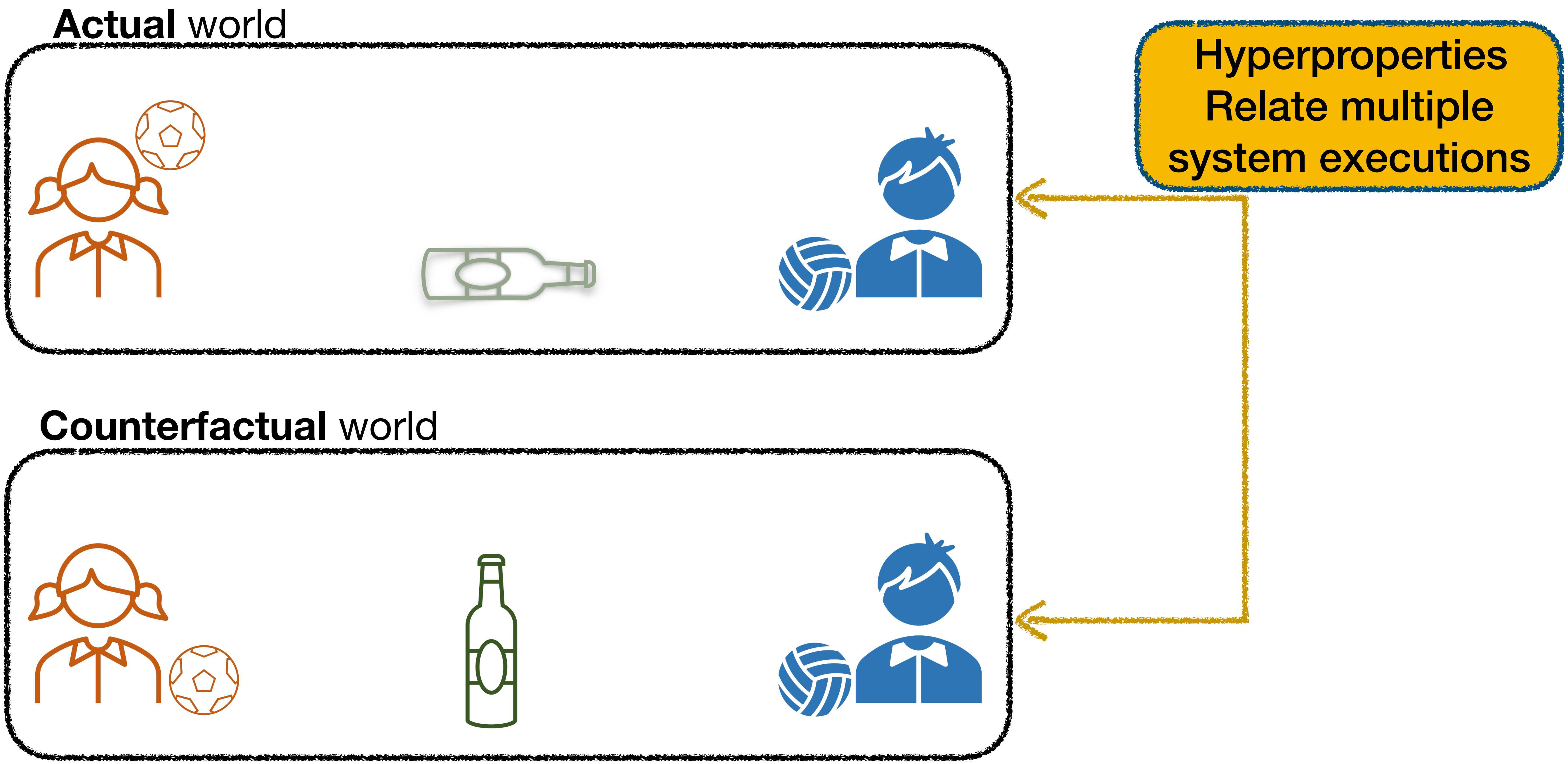


Actual Causality



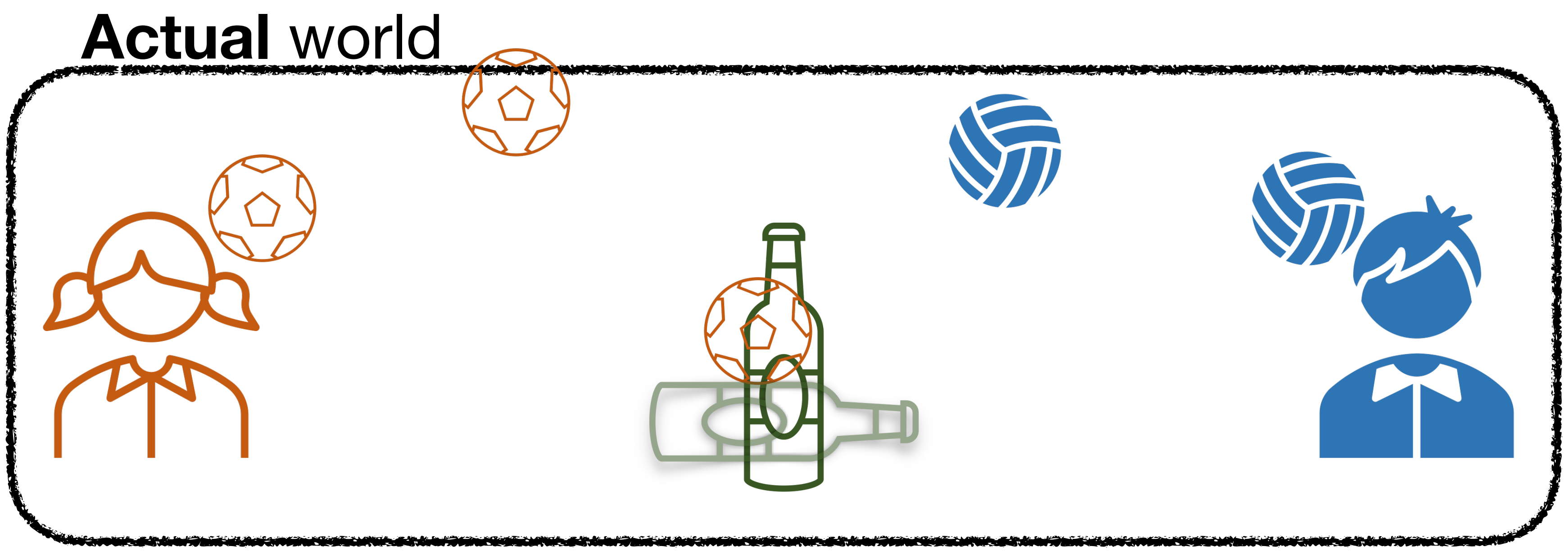
Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

Actual Causality



Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

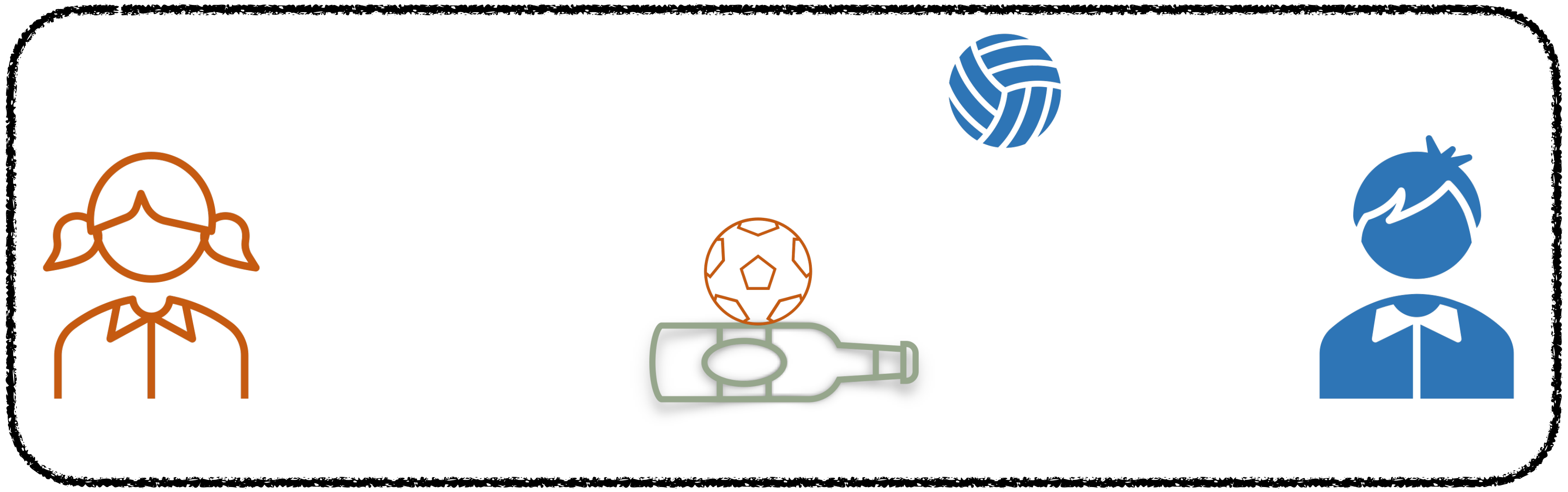
Actual Causality



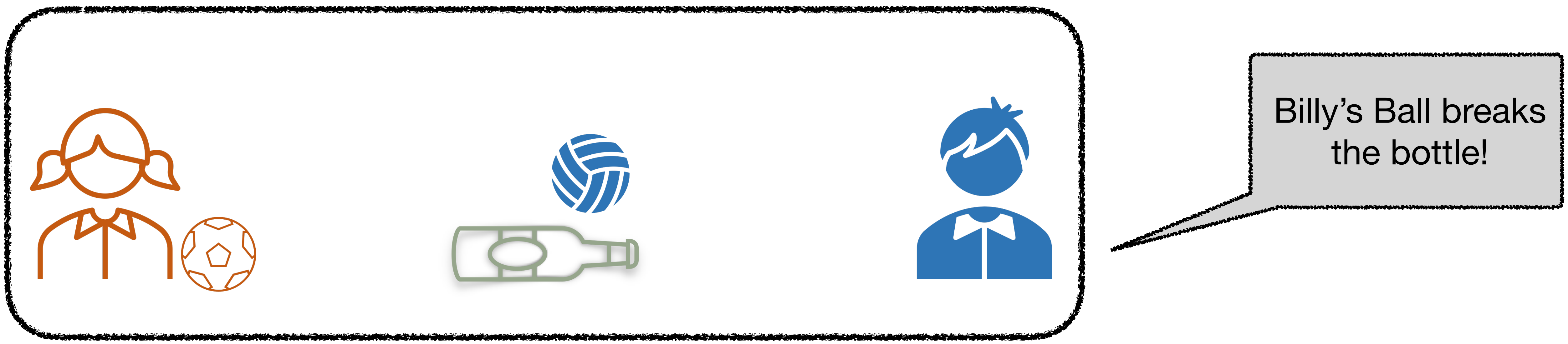
Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

Actual Causality

Actual world



Counterfactual world



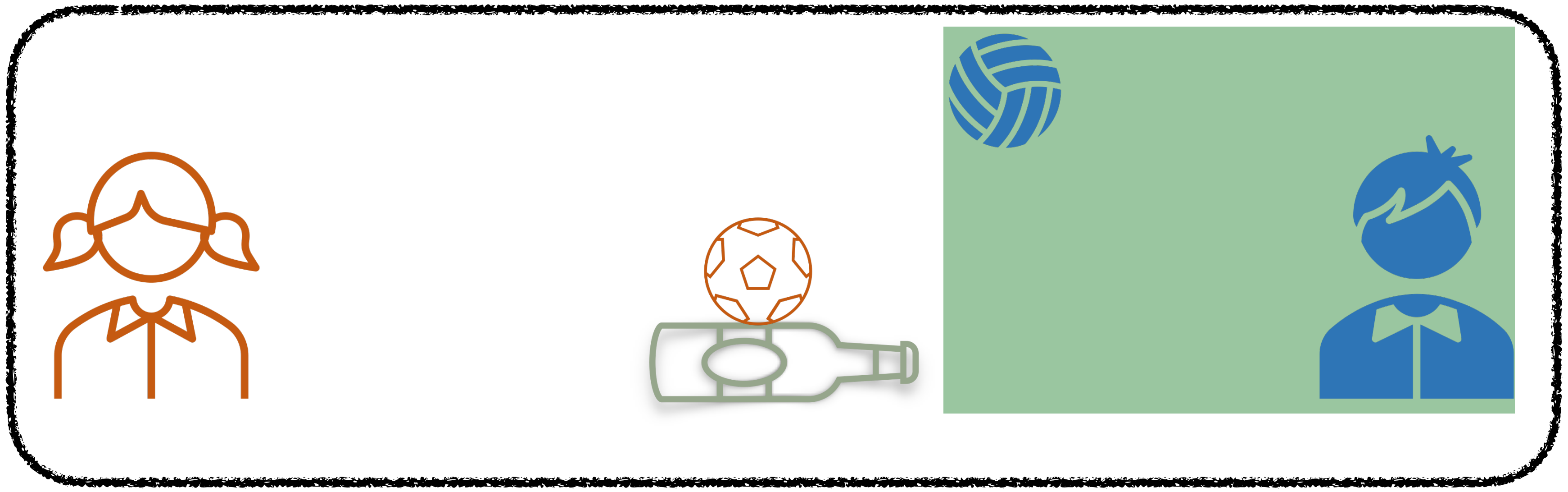
Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

Actual Causality

Contingencies

preemption of causes

Actual world



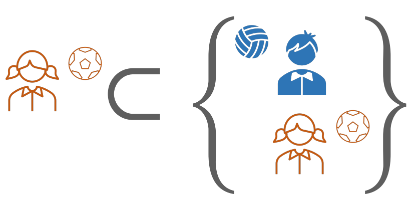
Counterfactual world + **contingency**

Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

Actual Causality

AC1: the cause appears in the actual world

AC2: for every counterfactual world there exists a contingency where effect does not hold



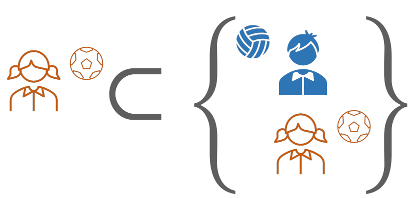
AC3: this is a minimal cause

Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

Actual Causality

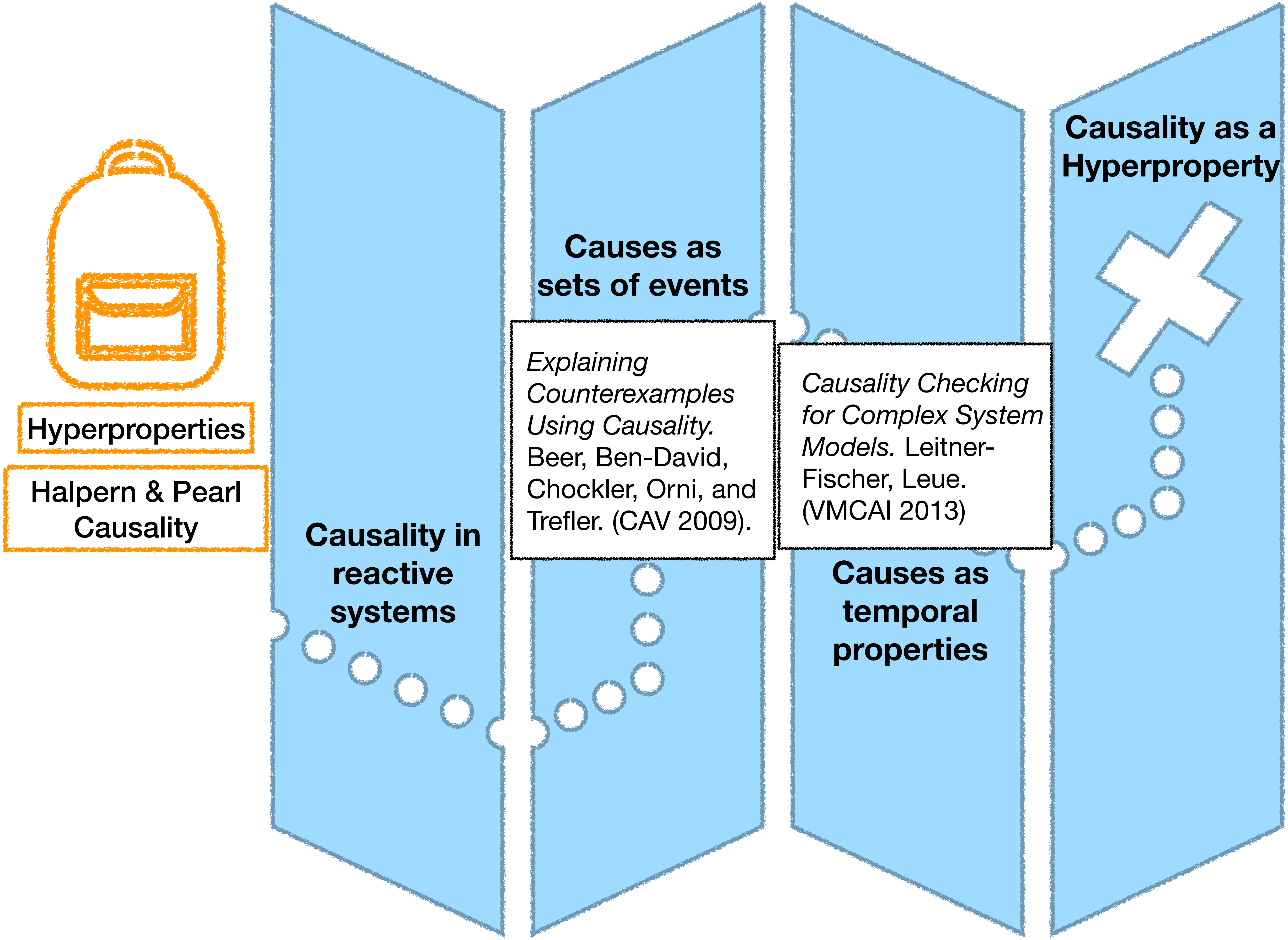
SAT: the cause appears in the actual world

CF: for every counterfactual world there exists a contingency where effect does not hold

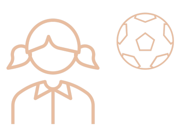


MIN: this is a minimal cause

Causes and Explanations: A Structural-Model Approach. Halpern and Pearl. Brit. J. Phil. Sci. 56 (2005).
A Modification of the Halpern-Pearl Definition of Causality. Halpern. (IJCAI 2015).

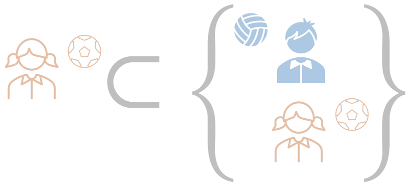

Hyperproperties
Halpern & Pearl Causality
Causality in reactive systems
Causes as sets of events
Explaining Counterexamples Using Causality. Beer, Ben-David, Chockler, Orni, and Trefler. (CAV 2009).
Causality Checking for Complex System Models. Leitner-Fischer, Leue. (VMCAI 2013)
Causes as temporal properties
Causality as a Hyperproperty

Actual Causality for Hyperproperties



SAT

CF : for every counterfactual world there exists a contingency where effect does not hold



MIN

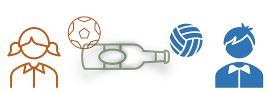
Actual Causality for Hyperproperties

∀* prefix:

$$\psi = \forall \pi_1 \forall \pi_2 \exists \pi_1' \exists \pi_2' . \varphi$$

$$\neg\psi = \exists \pi_1 \exists \pi_2 \forall \pi_1' \forall \pi_2' . \neg\varphi$$

- Effect: a violation of a **Hyperproperty** ψ



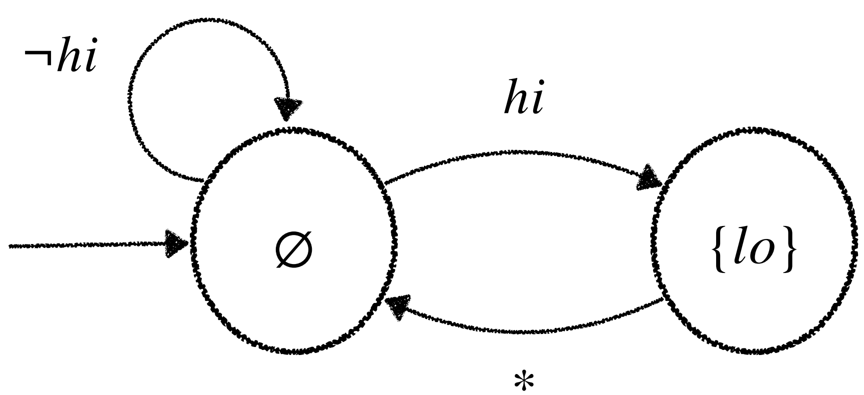
- Actual World: a set Γ of counterexample traces

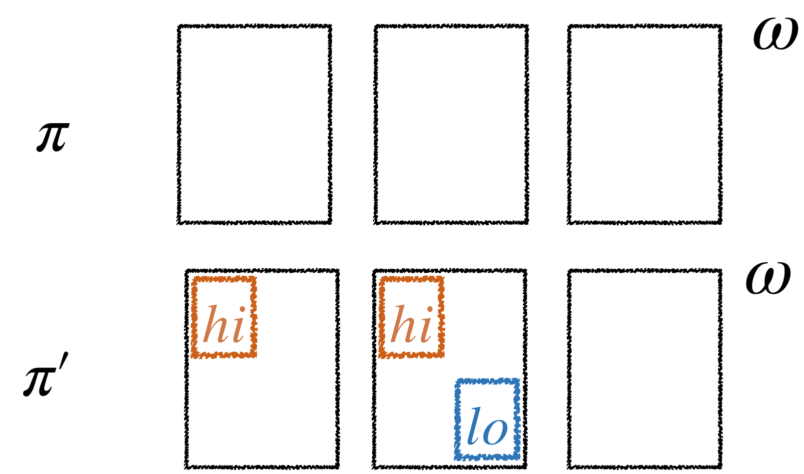
- Cause: set of events on the set of traces

Lasso-shaped

Explaining Hyperproperty Violations

$$\forall \pi \forall \pi' \, \square (li_\pi \leftrightarrow li_{\pi'}) \rightarrow \square (lo_\pi \leftrightarrow lo_{\pi'})$$



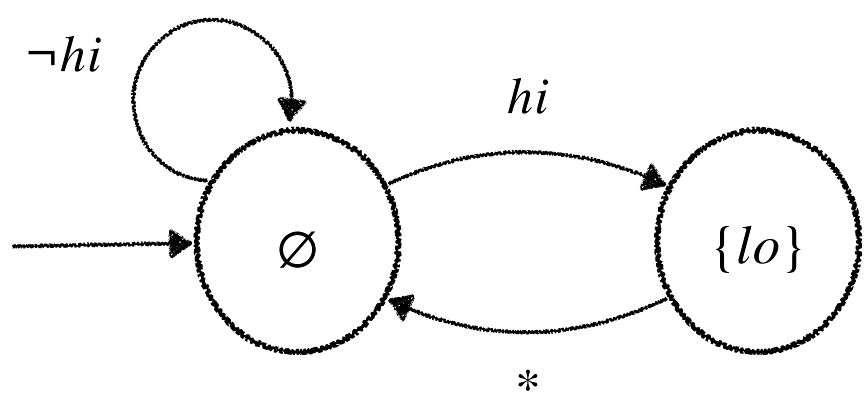


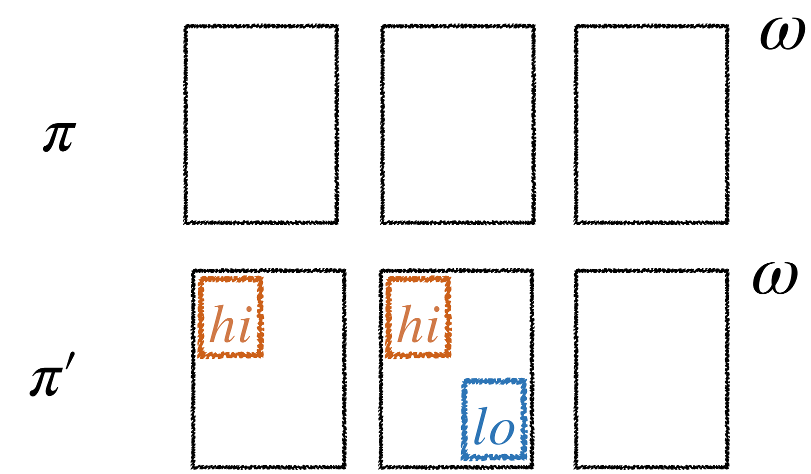
WHY?

Explaining Hyperproperty Violations

CF: ∀ counterfactual ∃ contingency s.t. φ holds

$$\forall \pi \forall \pi' \underbrace{\square (li_\pi \leftrightarrow li_{\pi'}) \rightarrow \square (lo_\pi \leftrightarrow lo_{\pi'})}_{\varphi}$$

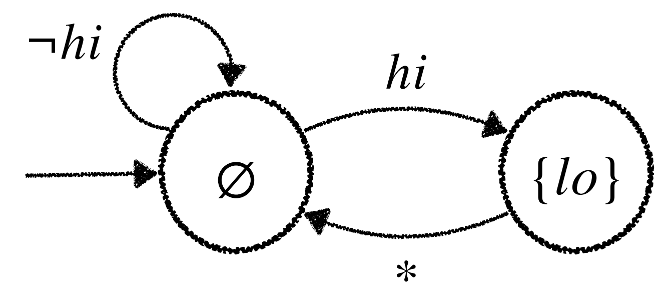


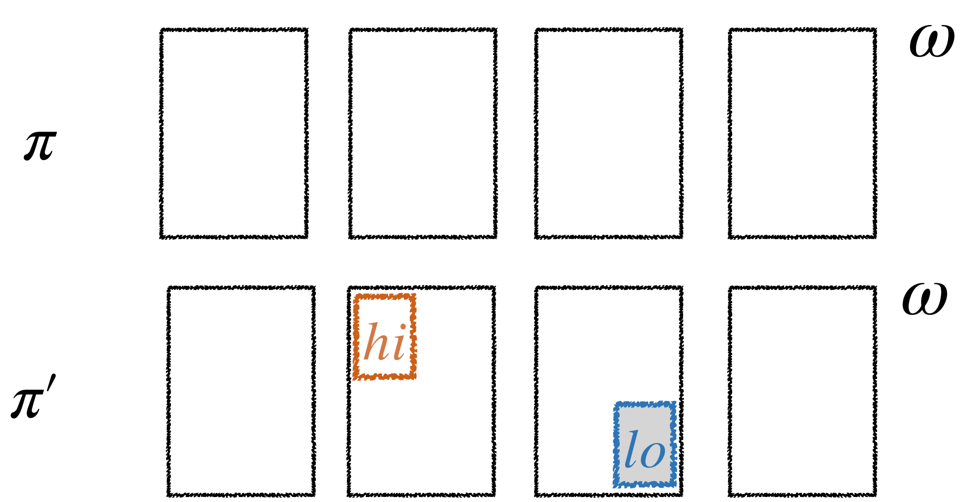


Explaining Hyperproperty Violations

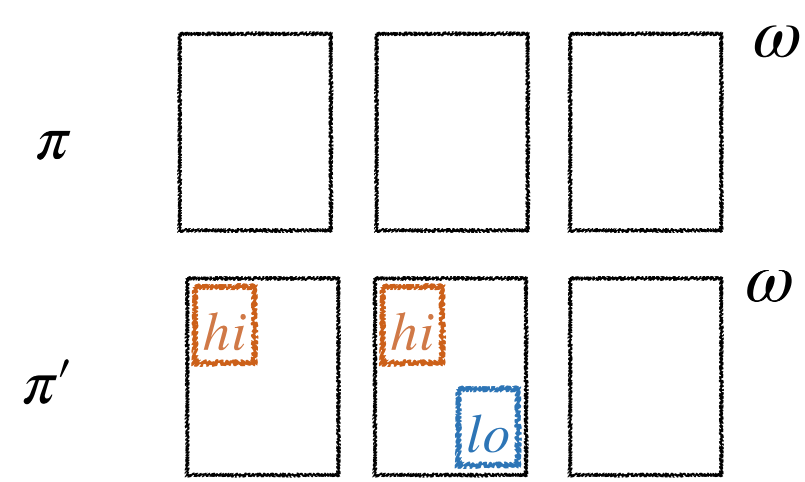
CF: ∀ counterfactual ∃ contingency s.t. φ holds

$$\forall \pi \forall \pi' \, \Box (li_{\pi} \leftrightarrow li_{\pi'}) \rightarrow \Box (lo_{\pi} \leftrightarrow lo_{\pi'})$$





Flip all events in C



$intervene(\Gamma, C, \varnothing)$

$$C = \{\langle hi, 0, \pi' \rangle\}$$

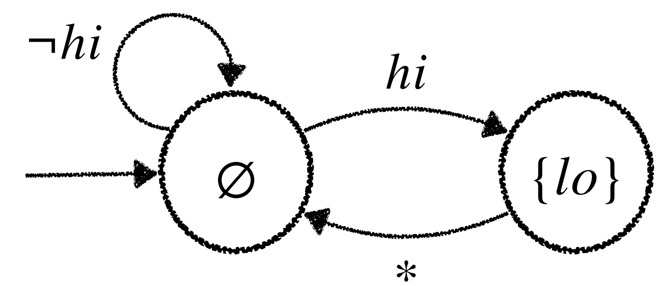
$$\Gamma = (\pi, \pi')$$

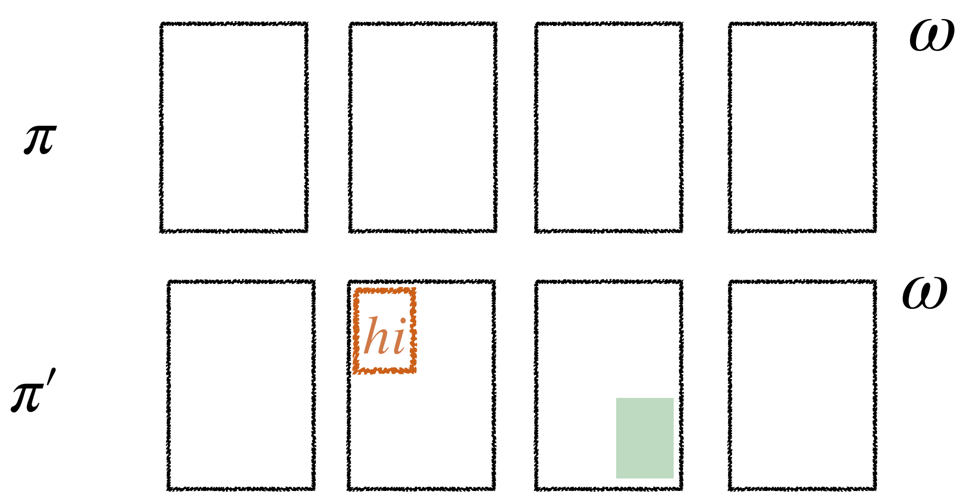
Cause - set of events

Explaining Hyperproperty Violations

CF: ∀ counterfactual ∃ contingency s.t. φ holds

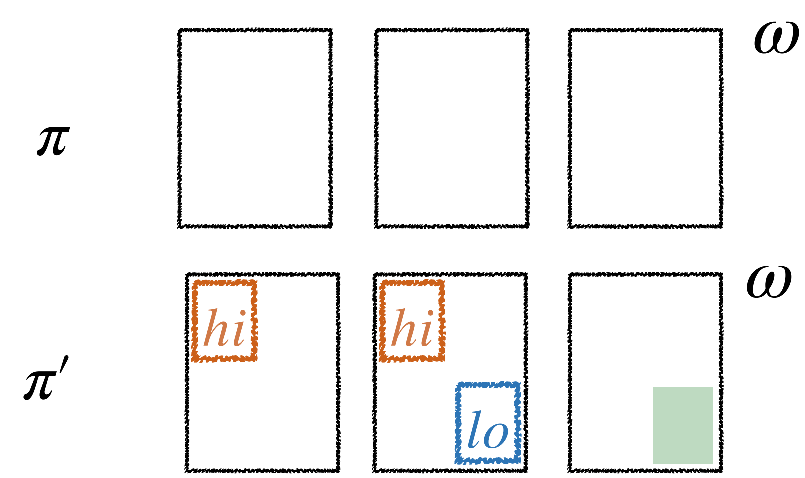
$$\forall \pi \forall \pi' \square (li_\pi \leftrightarrow li_{\pi'}) \rightarrow \square (lo_\pi \leftrightarrow lo_{\pi'})$$ ✓





Flip all events in C

Setting back to values of the original world



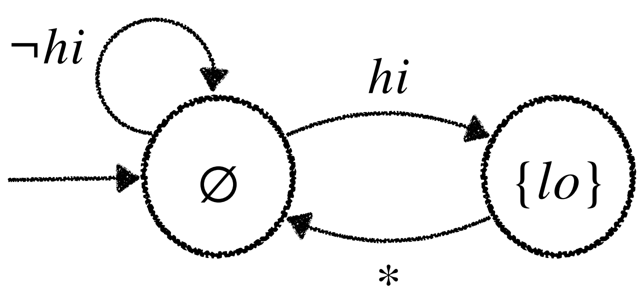
$intervene(\Gamma, C, \{\langle lo, 2, \pi' \rangle\})$

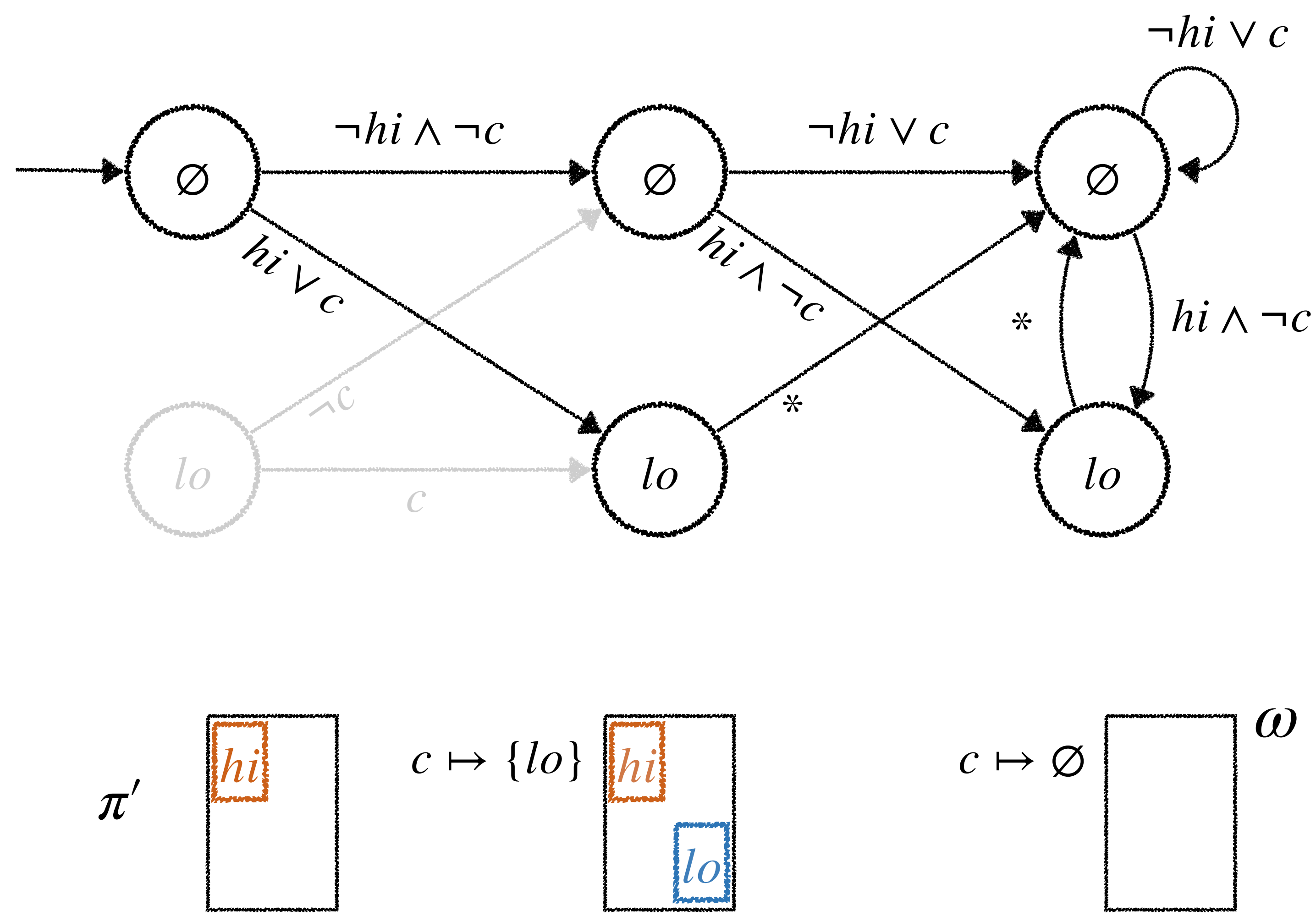
$C = \{\langle hi, 0, \pi' \rangle\}$

$\Gamma = (\pi, \pi')$

Cause - set of events

Computing Contingencies





an input c_o for each output o

Counterfactual automaton: additional inputs $[c]$ to set a contingency

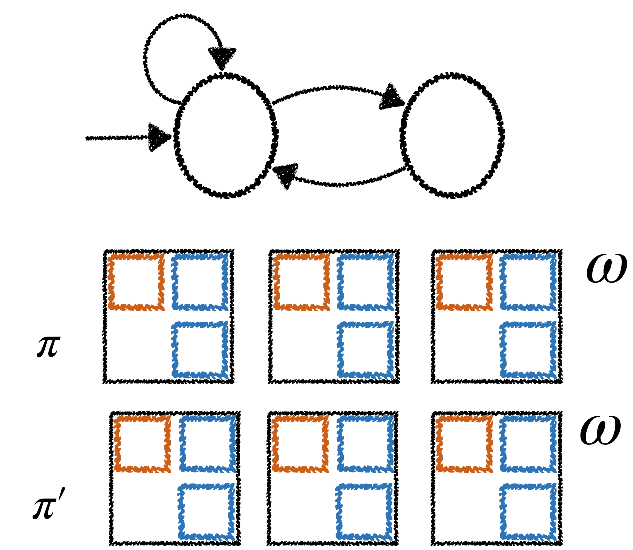
Actual Causality for Hyperproperties

Find C such that

SAT: $\Gamma \models C$

CF: $\forall$ counterfactual $\exists$ contingency s.t. φ holds

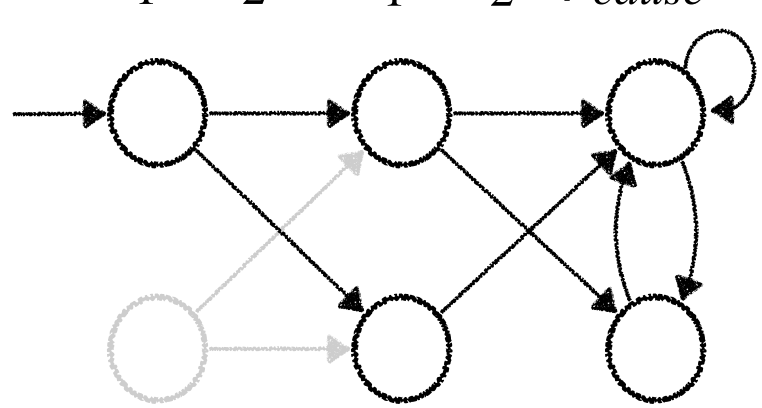
MIN: no subset of C satisfies SAT & CF



$$\forall \pi \forall \pi' \square (li_\pi \leftrightarrow li'_\pi) \rightarrow \square (lo_\pi \leftrightarrow lo'_\pi)$$

Finding a cause as a hyperproperty

$$\exists \pi_1 \exists \pi_2 . \forall \pi'_1 \forall \pi'_2 . \psi_{cause}$$

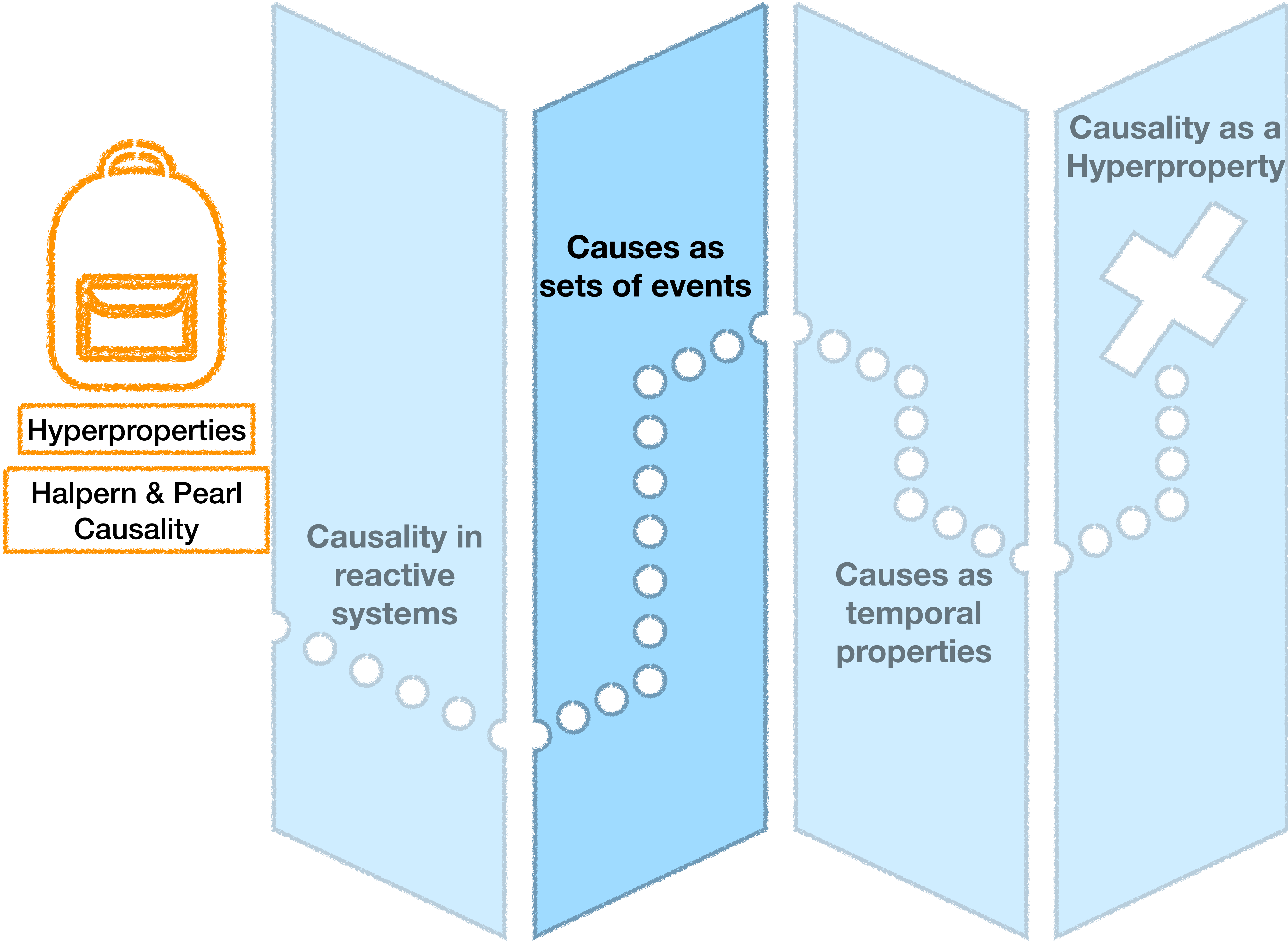


HyperLTL model checking

Events on π_1, π_2 correspond to the cause

π'_1, π'_2 represent other possible (not minimal) causes

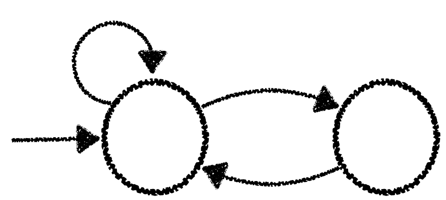
Algorithms for Model Checking HyperLTL and HyperCTL*. Finkbeiner, Rabe, Sánchez. (CAV 2015)

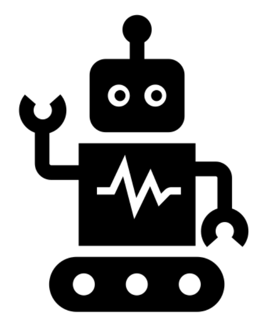

Hyperproperties
Halpern & Pearl Causality
Causality in reactive systems
Causes as sets of events
Causes as temporal properties
Causality as a Hyperproperty

Computing Actual Causes

largest candidate cause C — SAT dependencies

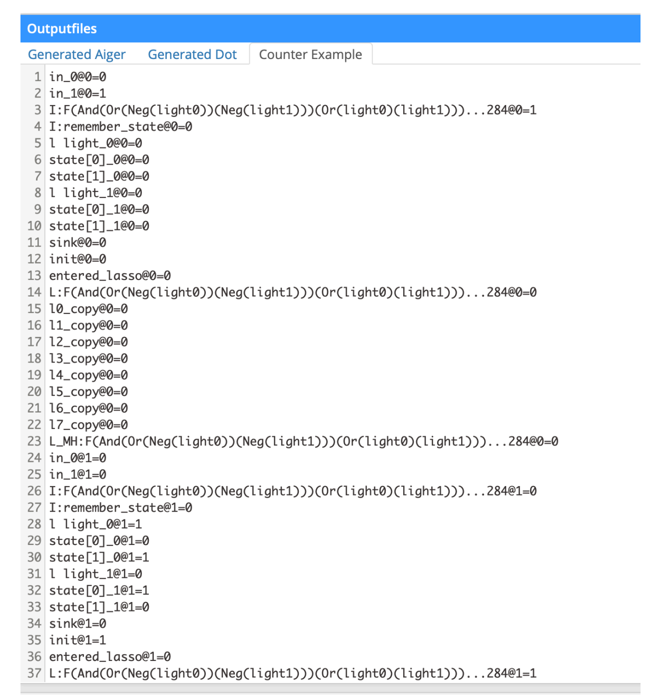
$\forall \pi_1, \ldots \forall \pi_n \,.\, \varphi$





MCHyper

$\Gamma = \pi_1, \ldots \pi_n$


Outputfiles

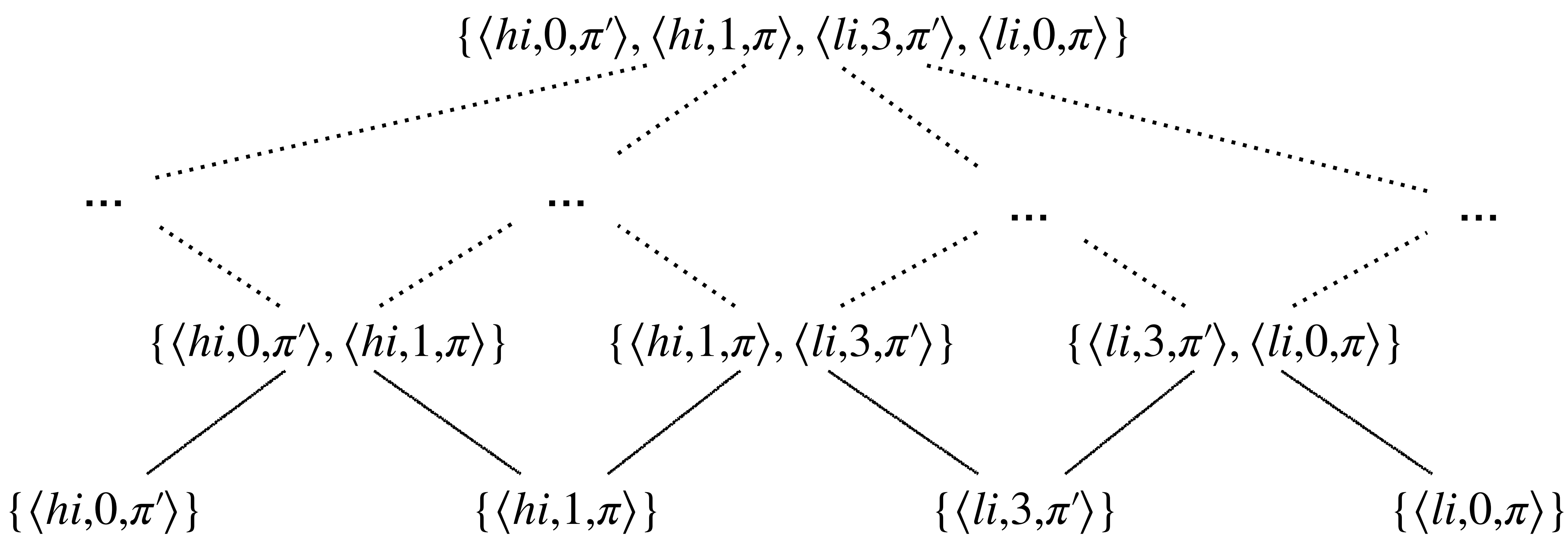
Generated Aiger | Generated Dot | Counter Example

```
in_0@0=0
in_1@0=1
I:F(And(Or(Neg(light0))(Neg(light1)))(Or(light0)(light1)))...284@0=1
I:remember_state@0=0
l light_0@0=0
state[0]_0@0=0
state[1]_0@0=0
l light_1@0=0
state[0]_1@0=0
state[1]_1@0=0
sink@0=0
init@0=0
entered_lasso@0=0
L:F(And(Or(Neg(light0))(Neg(light1)))(Or(light0)(light1)))...284@0=0
l0_copy@0=0
l1_copy@0=0
l2_copy@0=0
l3_copy@0=0
l4_copy@0=0
l5_copy@0=0
l6_copy@0=0
l7_copy@0=0
L_MH:F(And(Or(Neg(light0))(Neg(light1)))(Or(light0)(light1)))...284@0=0
in_0@1=0
in_1@1=0
I:F(And(Or(Neg(light0))(Neg(light1)))(Or(light0)(light1)))...284@1=0
I:remember_state@1=0
l light_0@1=1
state[0]_0@1=0
state[1]_0@1=1
l light_1@1=0
state[0]_1@1=1
state[1]_1@1=0
sink@1=0
init@1=1
entered_lasso@1=0
L:F(And(Or(Neg(light0))(Neg(light1)))(Or(light0)(light1)))...284@1=1
```

Algorithms for Model Checking HyperLTL and HyperCTL*. Finkbeiner, Rabe, Sánchez. (CAV 2015)

Computing Actual Causes

largest candidate cause C — SAT dependencies



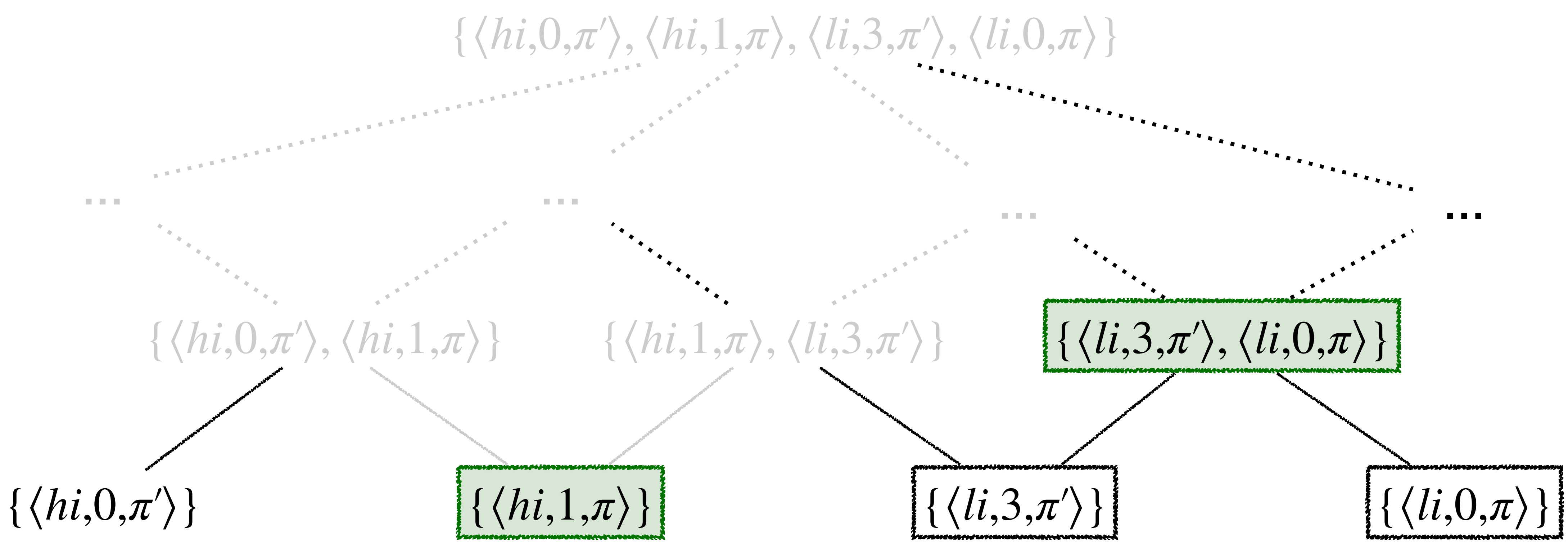
SAT: $\Gamma \models C$

CF: $\forall$ counterfactual $\exists$ contingency s.t. φ holds

MIN: no subset of C satisfies SAT & CF

Computing Actual Causes

largest candidate cause C — SAT dependencies



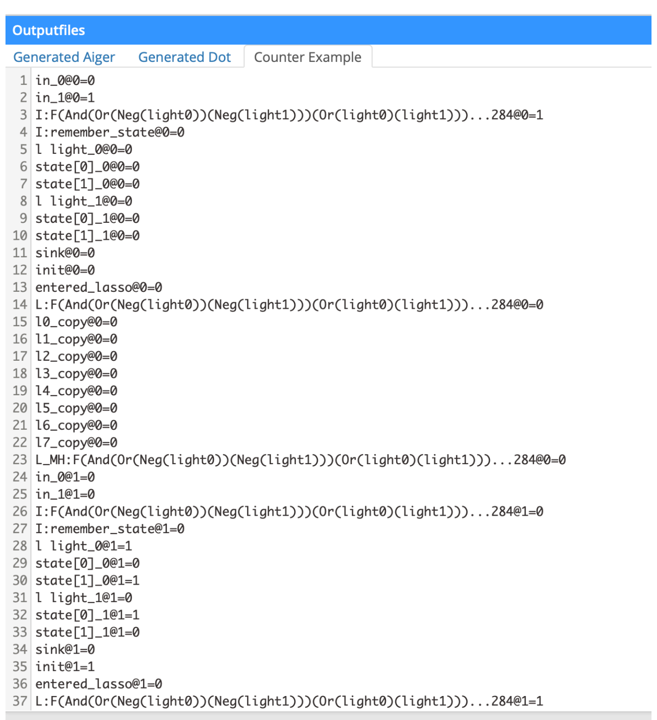
SAT: $\Gamma \models C$

CF: $\forall$ counterfactual $\exists$ contingency s.t. φ holds

MIN: no subset of C satisfies SAT & CF

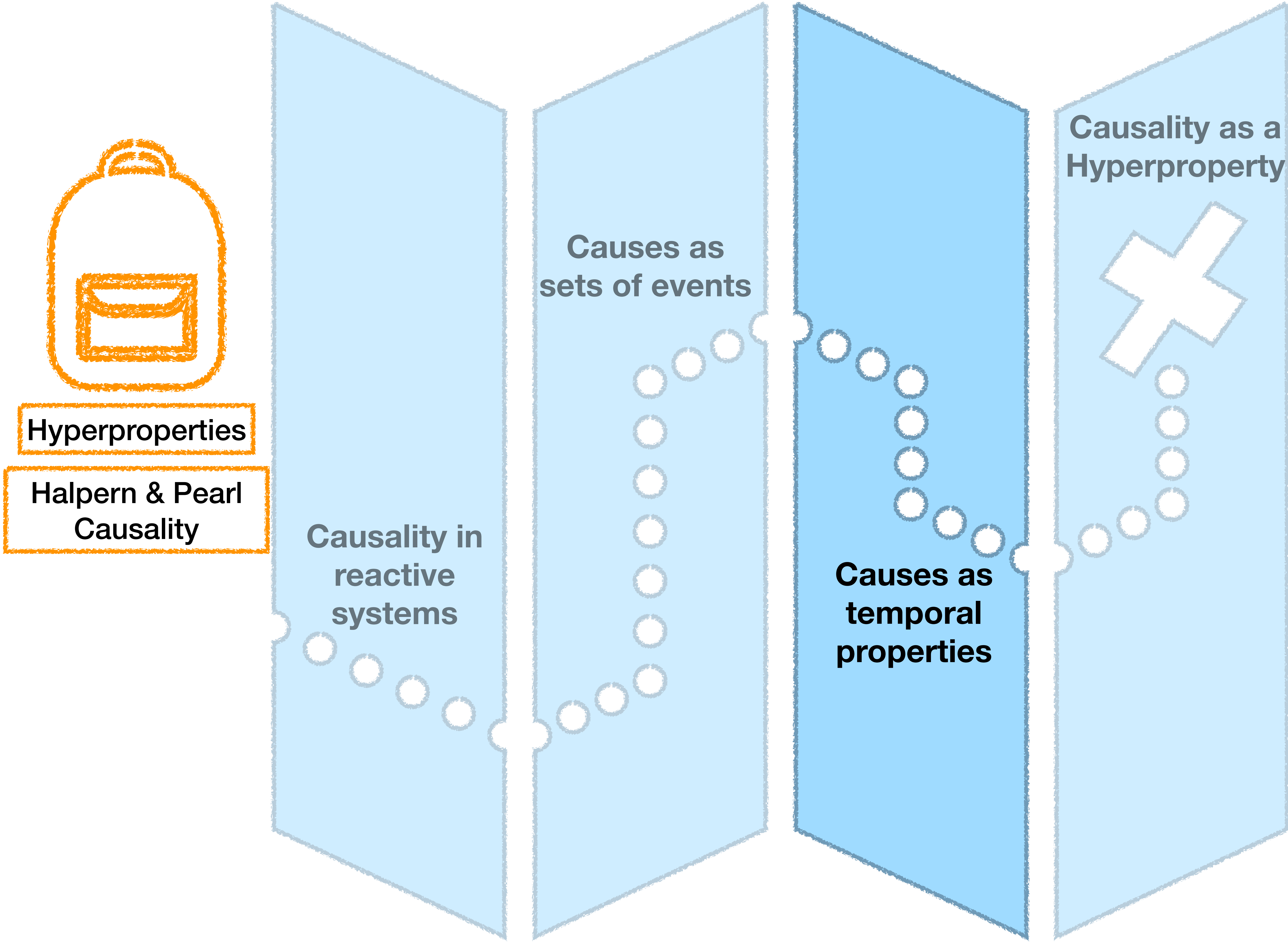
Experiments

Instance	$\lvert\Gamma\rvert$	$\lvert\varphi\rvert$	$\#(\mathcal{C})$	time(ms)
Running example (paper)	10	9	2	55
Security in & out	35	19	8	798
Drone example 1	24	19	5	367
Drone example 2	18	36	3	256
Asymmetric arbiter '19	28	35	10	490
Asymmetric arbiter	72	35	24	1480



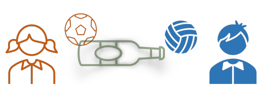
$$\{\langle hi, 0, \pi' \rangle\}$$

$$\{\langle \neg hi, 0, \pi \rangle\}$$


Hyperproperties
Halpern & Pearl Causality
Causality in reactive systems
Causes as sets of events
Causes as temporal properties
Causality as a Hyperproperty

Causes as Trace Properties

- Effect: a violation of an ω-regular property ψ



- Actual World: a counterexample trace

Lasso-shaped

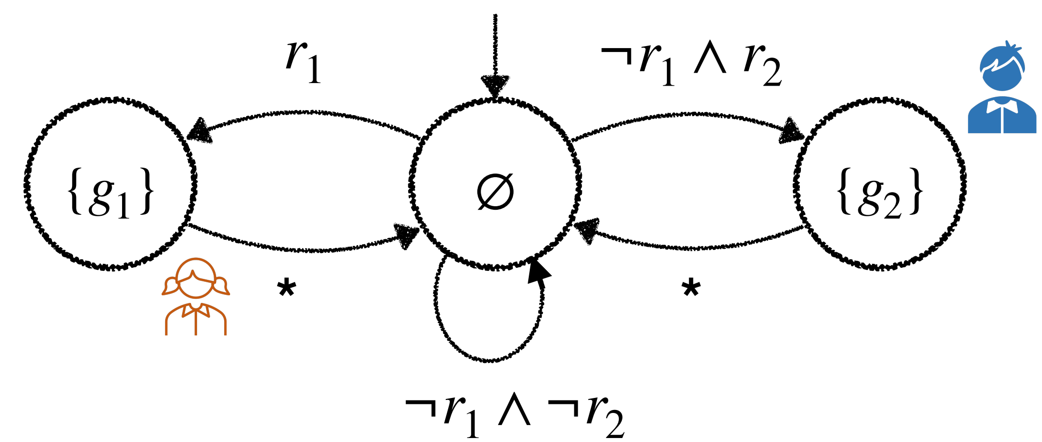
- Cause: an ω-regular property

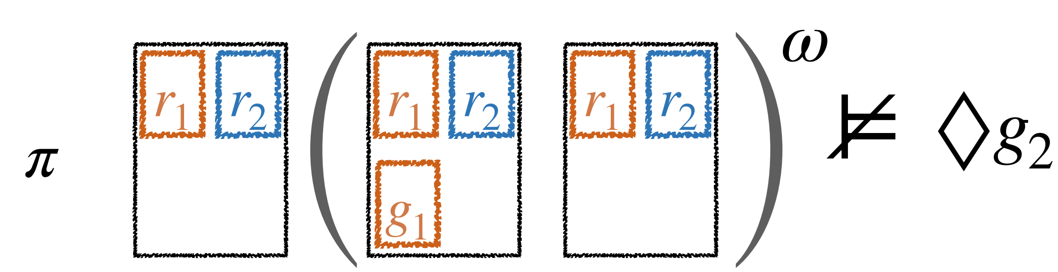
Quantified Propositional Temporal Logic — QPTL

LTL + quantification over propositions

$\exists q \,.\, q \wedge \Box (q \leftrightarrow \bigcirc \neg q) \wedge \Box (q \rightarrow a)$ — "a holds at every odd position"

Causes as Trace Properties

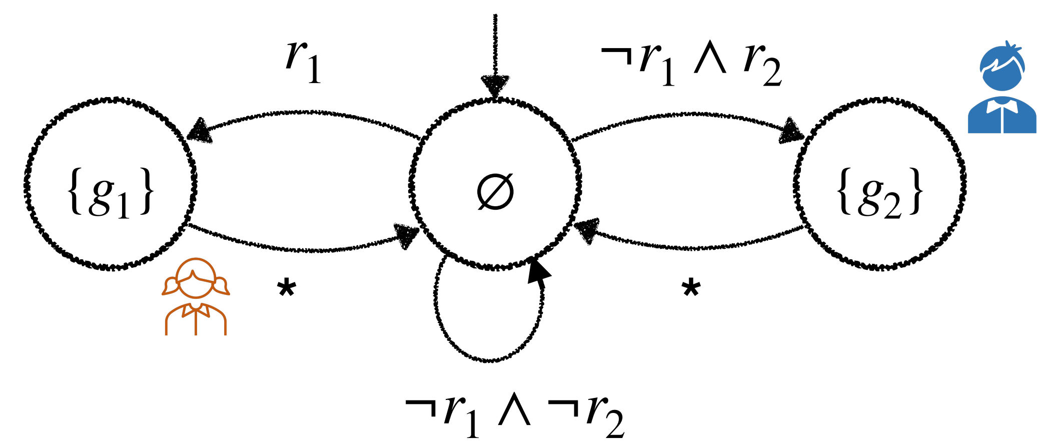




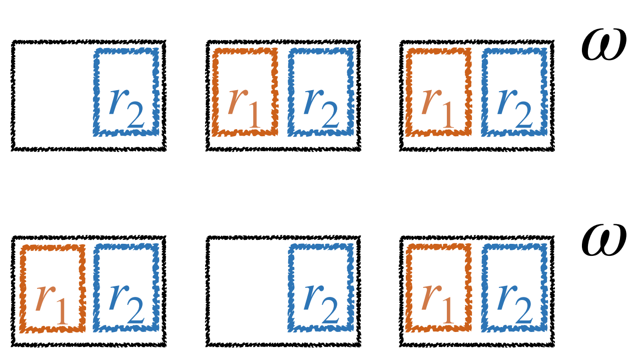
Causes as Trace Properties

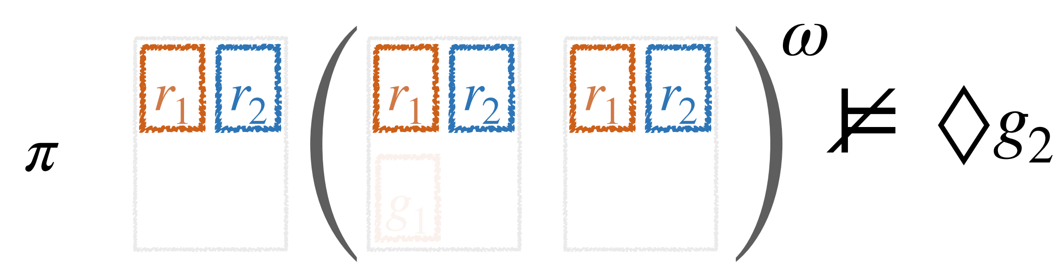
CF: ∀ counterfactual ∃ contingency s.t. $\Diamond g_2$ holds

Closest input sequences s.t. C doesn't hold



$$C = r_1 \wedge \bigcirc r_1$$
$$\neg C = \neg r_1 \vee \bigcirc \neg r_1$$





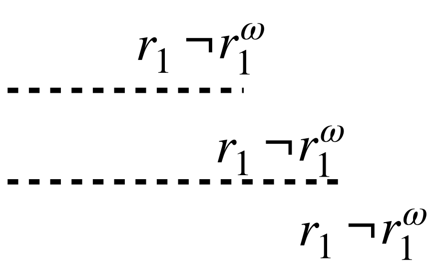
Causes as Trace Properties

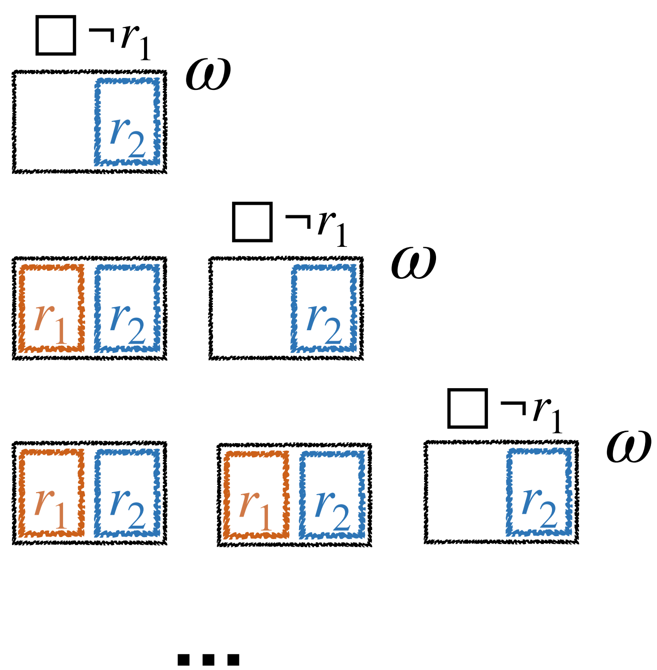
CF: ∀ counterfactual ∃ contingency s.t. $\Diamond g_2$ holds

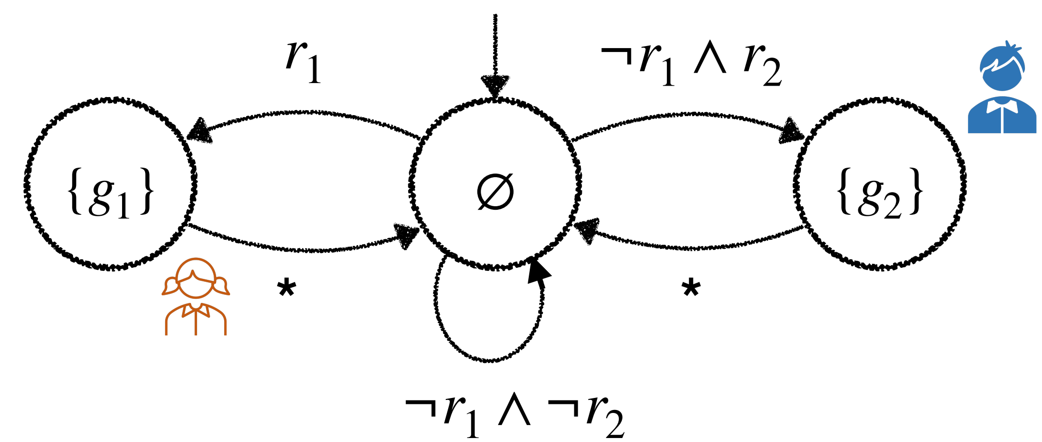
Closest input sequences s.t. C doesn't hold

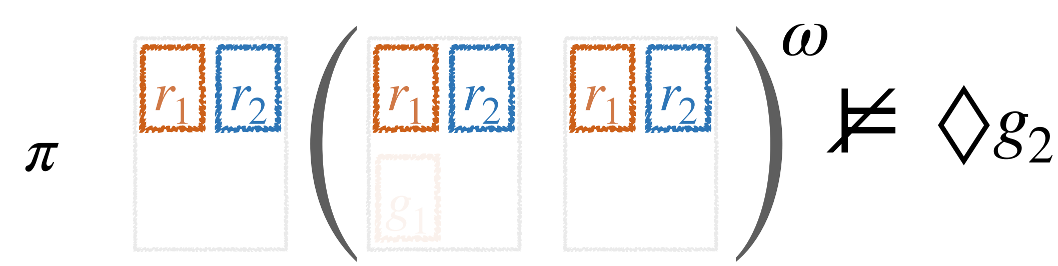
$$C = \Box \Diamond r_1$$

$$\neg C = \Diamond \Box \neg r_1$$







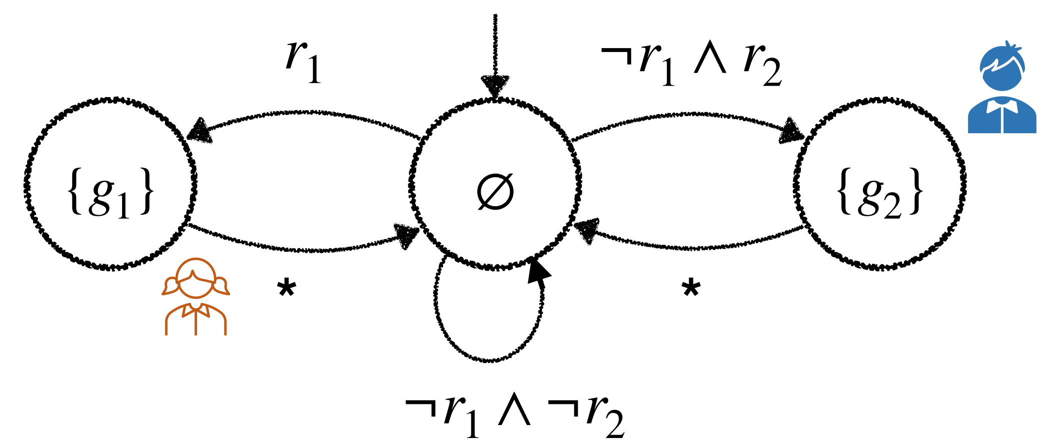


Compare traces that have the same **rejection structure**

Causes as Trace Properties

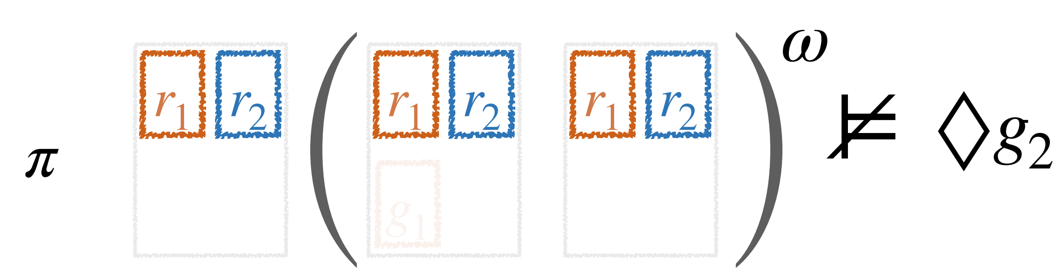
CF: ∀ counterfactual ∃ contingency s.t. $\Diamond g_2$ holds

Closest input sequences s.t. C doesn't hold



HyperQPTL formula

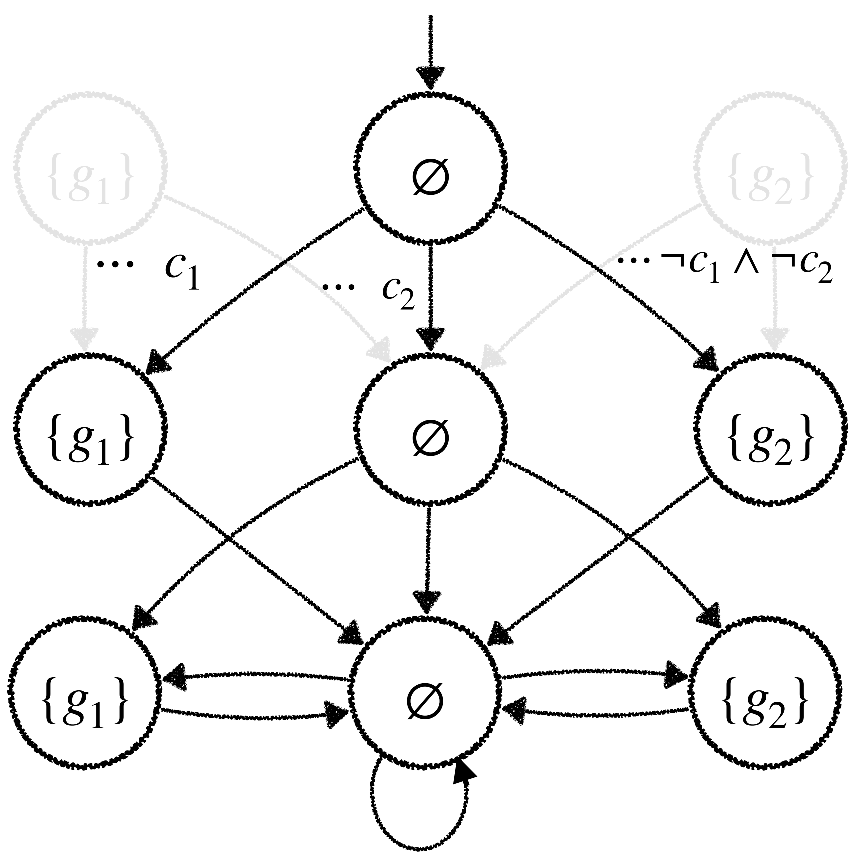
$\psi_{struct}(\pi_1, \pi_2)$: π_1, π_2 satisfy all sub-formulas of C at the same positions

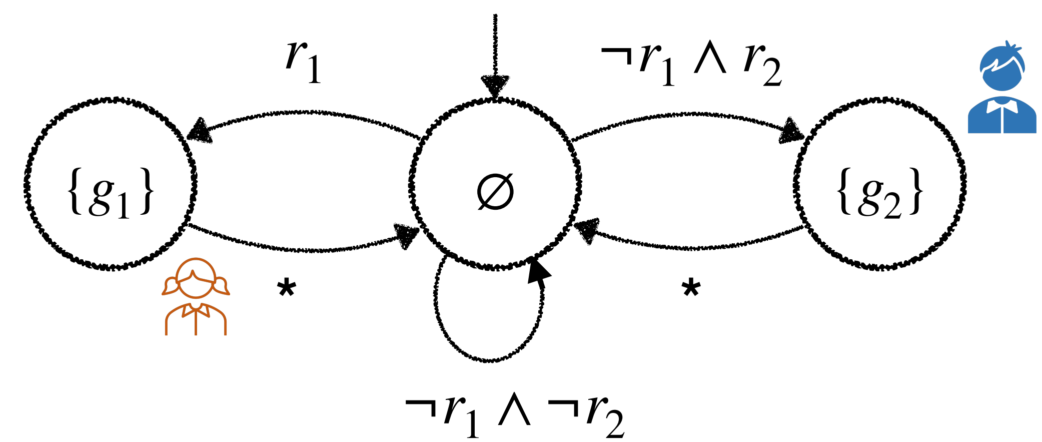


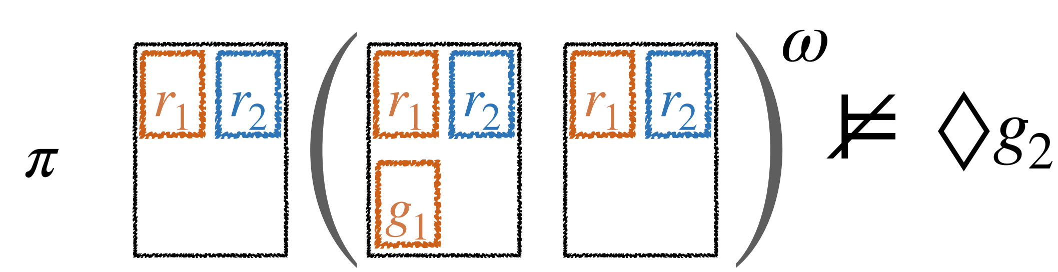
Compare traces that have the same **rejection structure**

Causes as Trace Properties

CF: ∀ counterfactual ∃ contingency s.t. $\Diamond g_2$ holds





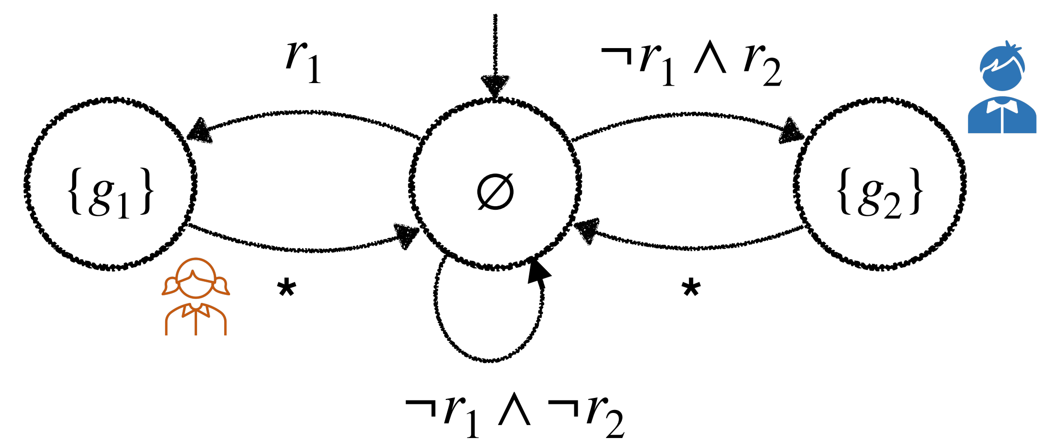


Counterfactual automaton additional inputs $[c_1, c_2]$ set a contingency

Causes as Trace Properties

MIN: There is no C' such that $C' \rightarrow C$

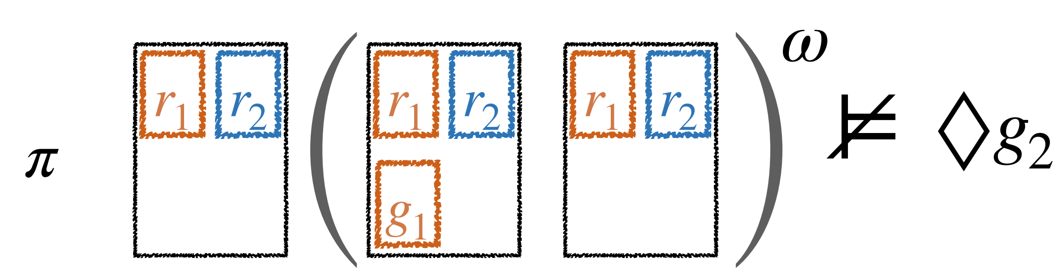
and $C' \vDash$ SAT & CF



HyperQPTL formula

No lasso-shaped trace can be removed from C

a trace that does not satisfy the effect, or does not contribute for counterfactual traces



Causes as Trace Properties

Given a candidate cause C, verify:

SAT: $\pi \vDash C$

CF: $\forall$ counterfactual $\exists$ contingency s.t. φ holds

MIN: There is no C' such that $C' \rightarrow C$ and $C' \vDash$ SAT & CF

$\Rightarrow$ C **is a cause of** φ **on** π

HyperQPTL

SAT: Verify C on π

CF: Counterfactuals: traces with the same rejection structure
Contingencies: using the counterfactual automaton

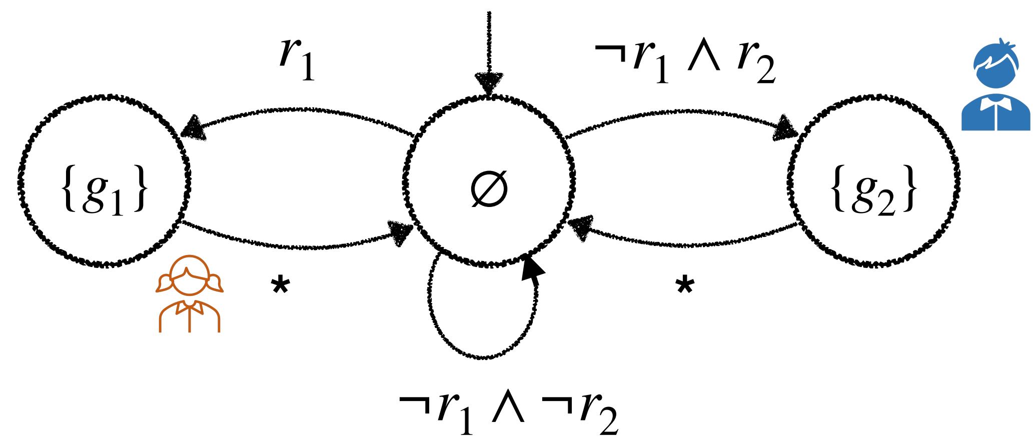
MIN: No lasso-shaped trace can be removed from C

Decidable via HyperQPTL model-checking!

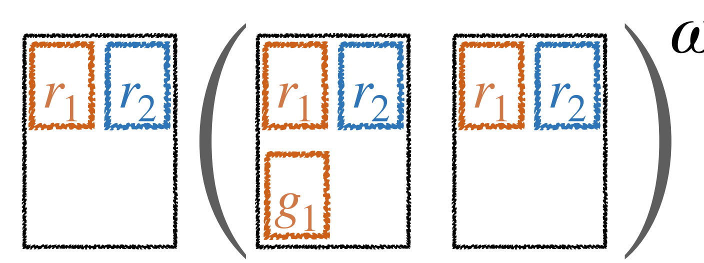
Rabe. A temporal logic approach to information-flow control. Ph.D. thesis (2016)

Unfair Arbiter

Is $\square\, r_1$ the cause for $\square \neg g_2$?



π



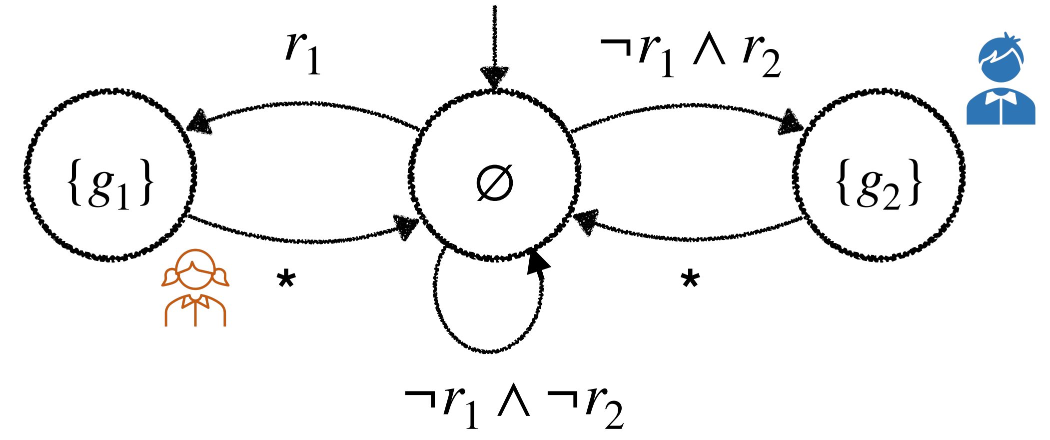
$\not\models \Diamond g_2$

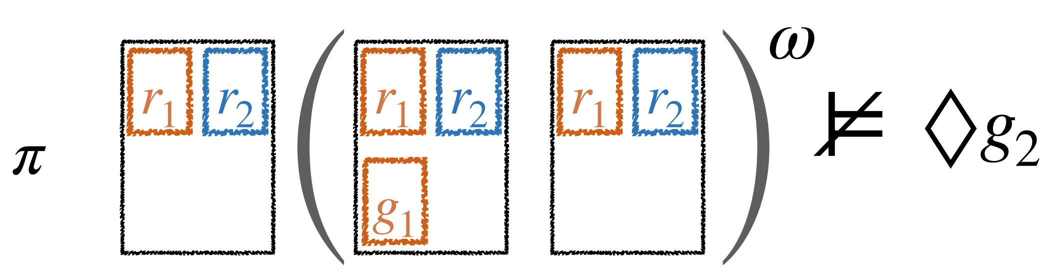
Unfair Arbiter

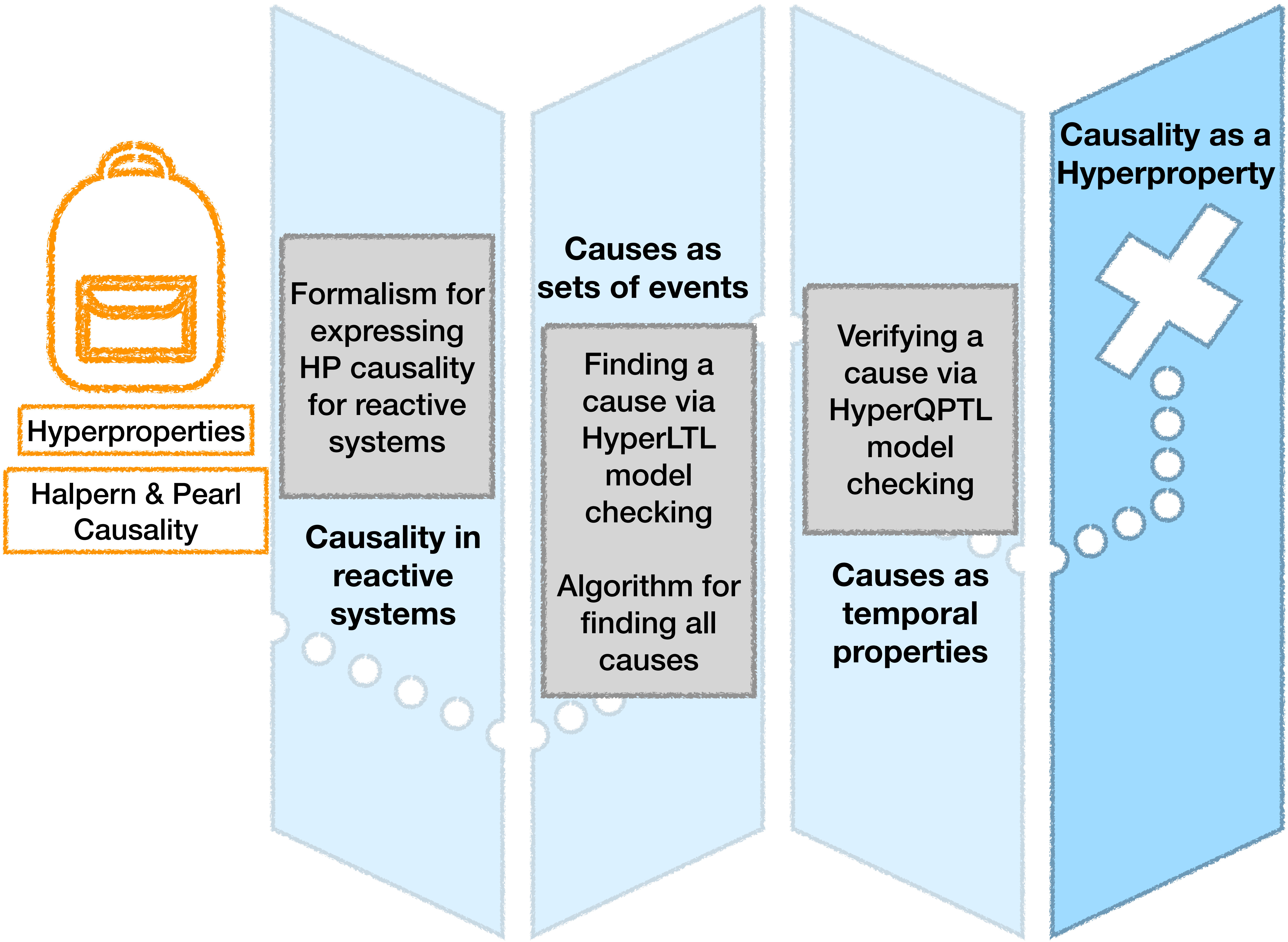
$$\exists q \,.\, q \wedge \Box (q \leftrightarrow \bigcirc \neg q) \wedge \Box (q \rightarrow r_1)$$

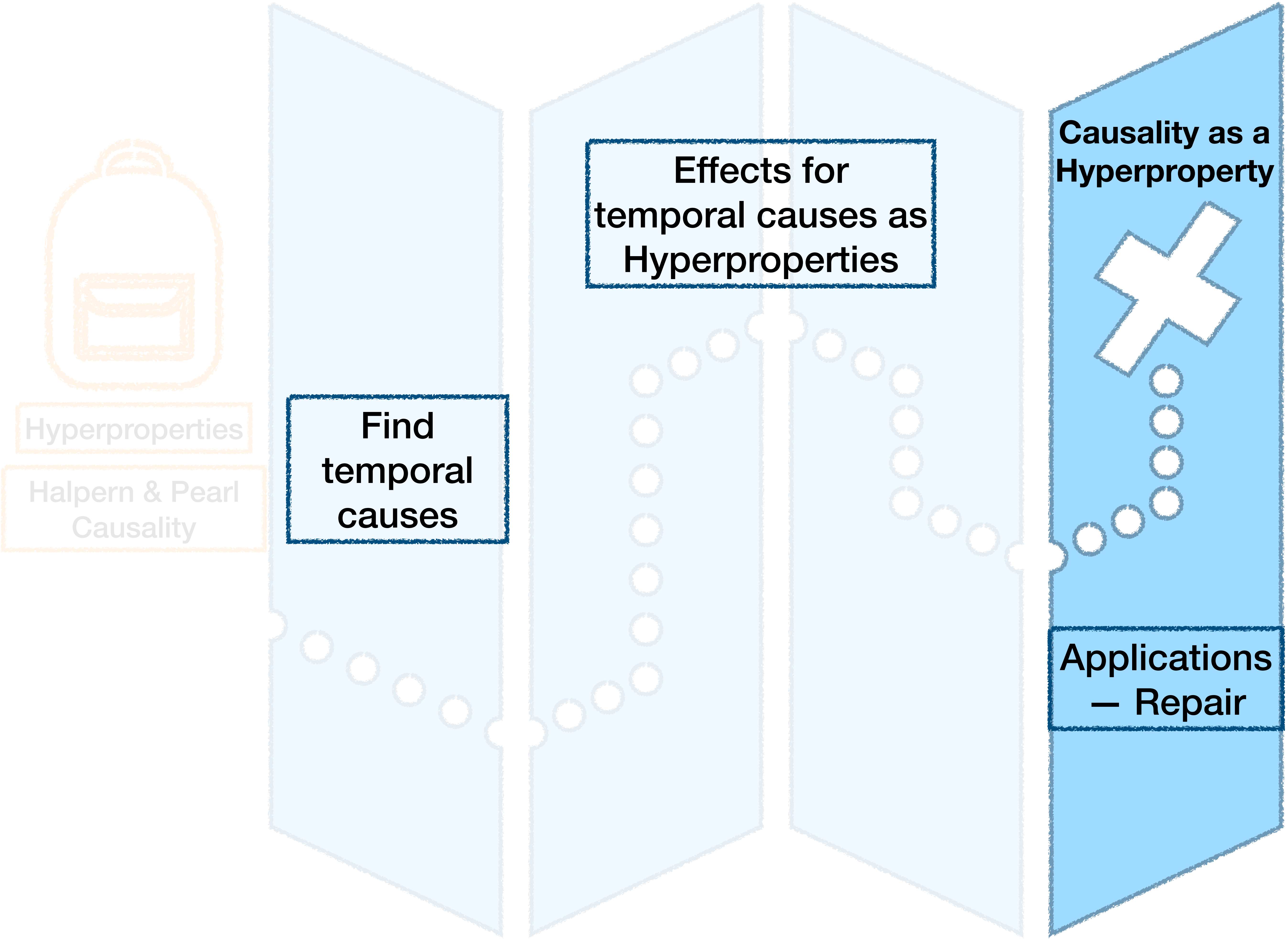
r_1 holds at every odd position

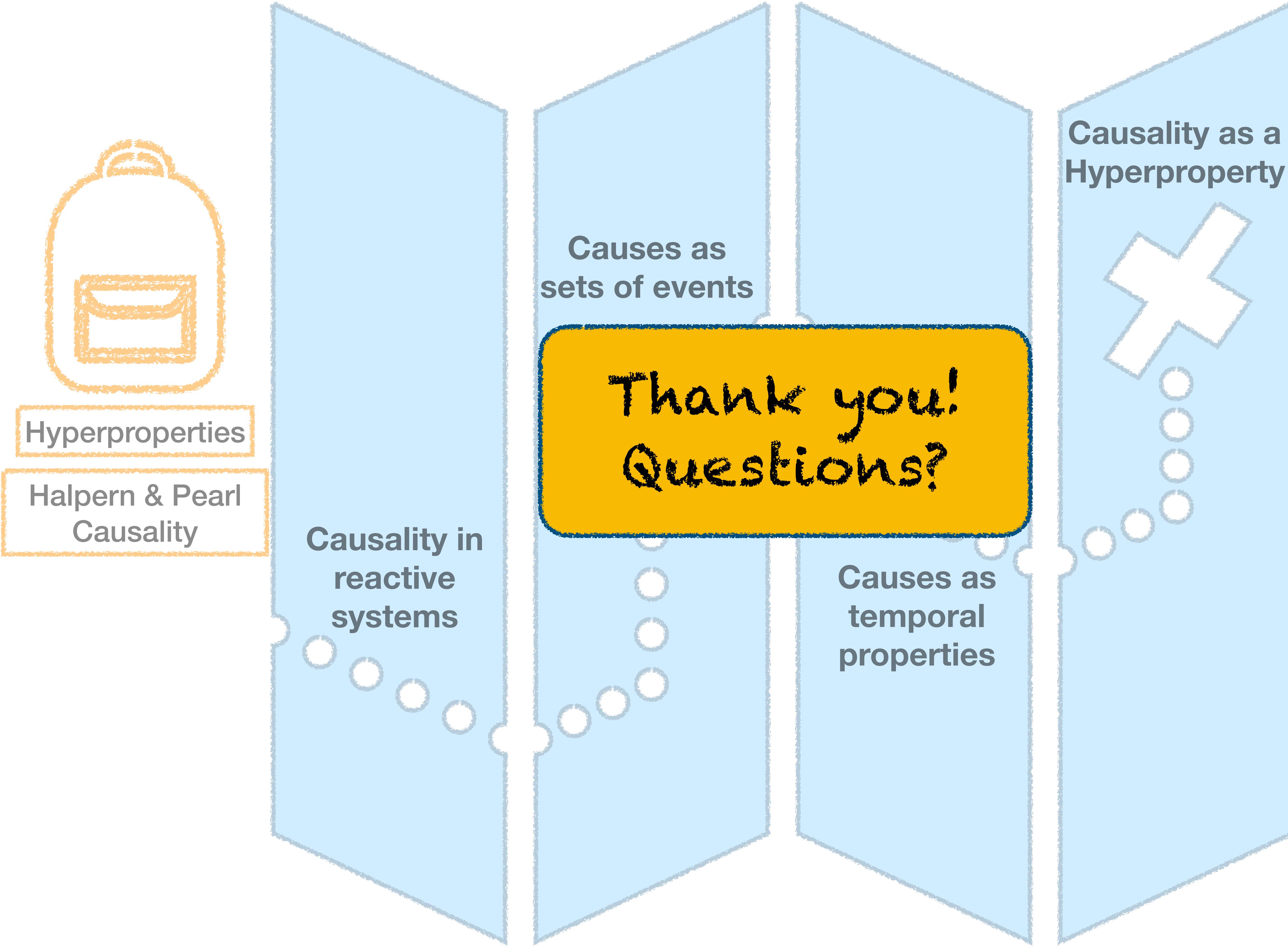
is a cause for $\Box \neg g_2$ on π






Hyperproperties
Halpern & Pearl Causality
Formalism for expressing HP causality for reactive systems
Causality in reactive systems
Causes as sets of events
Finding a cause via HyperLTL model checking
Algorithm for finding all causes
Verifying a cause via HyperQPTL model checking
Causes as temporal properties
Causality as a Hyperproperty


Hyperproperties
Halpern & Pearl Causality
Find temporal causes
Effects for temporal causes as Hyperproperties
Causality as a Hyperproperty
Applications — Repair


Hyperproperties
Halpern & Pearl Causality
Causality in reactive systems
Causes as sets of events
Thank you!
Questions?
Causes as temporal properties
Causality as a Hyperproperty