Temporal Causality in Reactive Systems

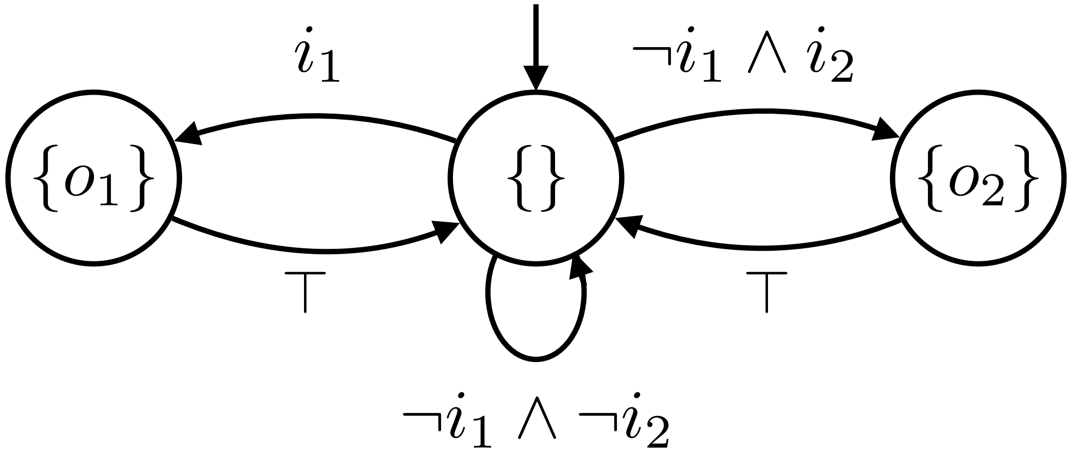
Norine Coenen[1], Bernd Finkbeiner[1], Hadar Frenkel[1],
Christopher Hahn[2], Niklas Metzger[1], and Julian Siber[1]

[1]*CISPA Helmholtz Center for Information Security*
[2]*Stanford University*

ATVA 2023 | 27 October 2023 | In Proceedings of ATVA 2022

Temporal Properties as Causes

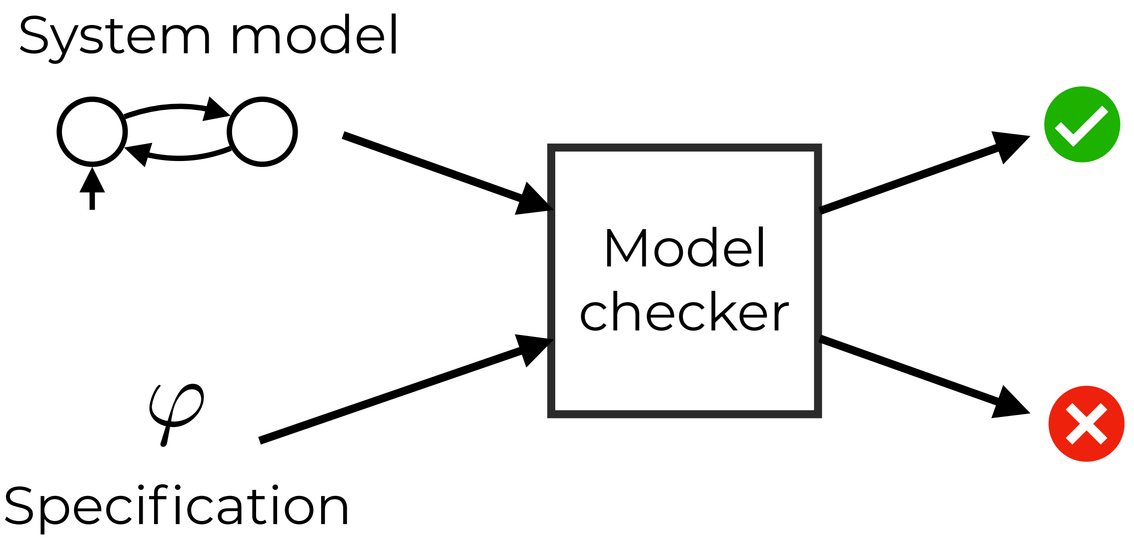
System:



Trace: $(\{i_1, i_2\}\{i_1, i_2, o_1\})^\omega$

Does $\Diamond i_1$ *cause* $\Box \neg o_2$?

Causes as Explanations for Model Checking



Counterexample $\{a,b\}\{a\}(\{a,b\})^\omega$

The *causes* for $\neg\varphi$ can explain the counterexample.

One solution: Highlighting

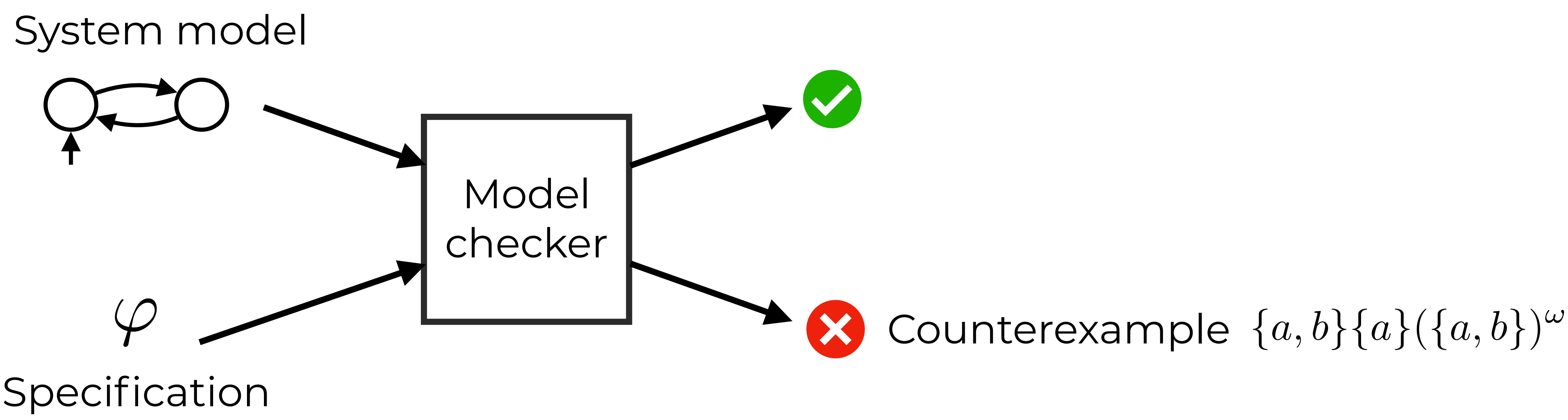
$\{a,\boxed{b}\}\{a\}(\{a,\boxed{b}\})^\omega$

$\Box\Diamond b$?

$\Diamond b$?

$b \wedge \bigcirc\bigcirc b$?

Causes as Explanations for Model Checking



The *causes* for $\neg\varphi$ can explain the counterexample.

One solution: Highlighting

$\{a,b\}\{a\}(\{a,b\})^\omega$

Our Solution: Property causes

$\Diamond b$ $\qquad$ $\Box\Diamond b$

$b \wedge \bigcirc\bigcirc b$

Outline

Define causality for trace properties

Infinitely many variables

Algorithm for checking causality

Infinitely many traces

Actual Causality[1,2]

For finite sets of events $(a, n) \in AP \times \mathbb{N}$ (proposition and time-point).

SAT: Cause and Effect have happened.

CF: If the cause had not happened (but everything else stayed the same), the effect would not have happened either.

MIN: There is no subset that satisfies the above.

[1] *Causes and Explanations: A Structural-Model Approach*. Halpern and Pearl (2005).
[2] *A Modification of the Halpern-Pearl Definition of Causality*. Halpern (2015).

C is a Cause for E iff...

SAT: π satisfies C and E .

CF: If the cause had not happened (but everything else stayed the same), the effect would not have happened either.

MIN: There is no smaller cause candidate that satisfies the above.

Distance Metrics

An adaption of similarity relations[3,4].

$$\pi = (\{i_1, i_2\}\{i_1, i_2, o_1\})^\omega$$

A *distance metric* $<_\pi^C$ orders traces w.r.t. their similarity to π .

$$\pi_1 <_\pi^C \pi_2 \quad \text{iff} \quad \mathit{zip}(\pi, \pi_1, \pi_2) \models$$

$$\square \bigwedge_{i \in I} \left((i_\pi \not\leftrightarrow i_{\pi_1}) \rightarrow (i_\pi \not\leftrightarrow i_{\pi_2}) \right) \wedge \Diamond \bigvee_{i \in I} (i_{\pi_1} \not\leftrightarrow i_{\pi_2})$$

$\Longrightarrow$ Causality is a *hyperproperty*.

---

[3] *A Theory of Counterfactuals*. Stalnaker (1968).
[4] *Counterfactuals*. Lewis (1973).

Counterfactual Input Sequences

$$\pi = (\{i_1, i_2\}\{i_1, i_2, o_1\})^\omega$$

The counterfactual input sequences are the closest sequences that negate C:

$$C_1 = i_1 \lor \bigcirc i_1 \quad \Rightarrow \quad \sigma_{\neg C_1} = \{i_2\}\{i_2\}\{i_1, i_2\}^\omega$$

and not, e.g.: $\{i_2\}\{i_2\}\{i_2\}^\omega$

$$C_2 = i_1 \land \bigcirc i_1 \quad \Rightarrow \quad \sigma^1_{\neg C_2} = \{i_2\}\{i_1, i_2\}\{i_1, i_2\}^\omega$$

$$\sigma^2_{\neg C_2} = \{i_1, i_2\}\{i_2\}\{i_1, i_2\}^\omega$$

Limit Assumption

$$\pi = (\{i_1, i_2\}\{i_1, i_2, o_1\})^\omega$$

The naive distance metric could be vacuously satisfied:

$$C_3 = \Box\Diamond i_1 \quad \Rightarrow \quad \begin{aligned} \sigma^1_{\neg C_3} &= \{i_2\}^\omega \\ \sigma^2_{\neg C_3} &= \{i_1, i_2\}\{i_2\}^\omega \\ &\dots \\ \sigma^k_{\neg C_3} &= \{i_1, i_2\}^k\{i_2\}^\omega \\ &\dots \end{aligned}$$

is closer than

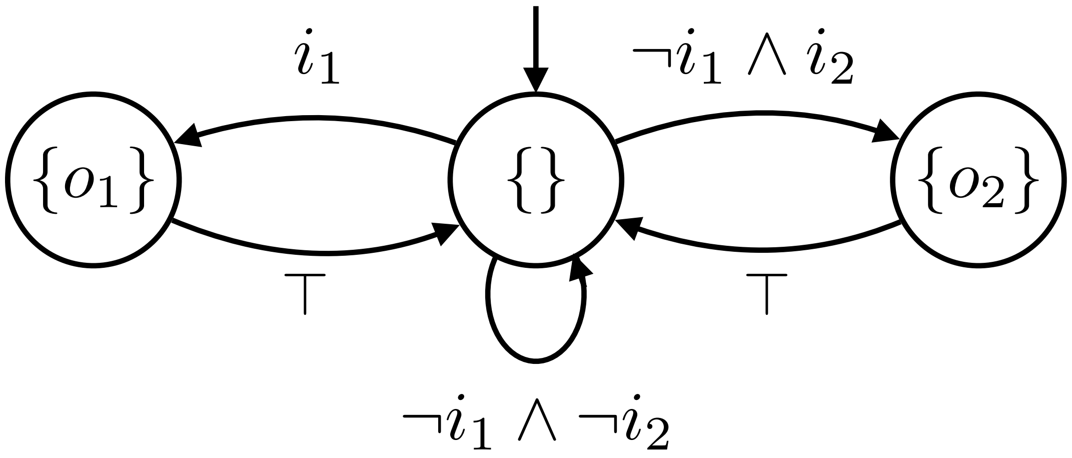
is closer than

We propose an extension that satisfies the *limit assumption.*

(See our recent work[5] on how to accommodate more general metrics).

[5] *Counterfactuals Modulo Temporal Logics*. Finkbeiner and Siber (2023).

Contingencies on Traces



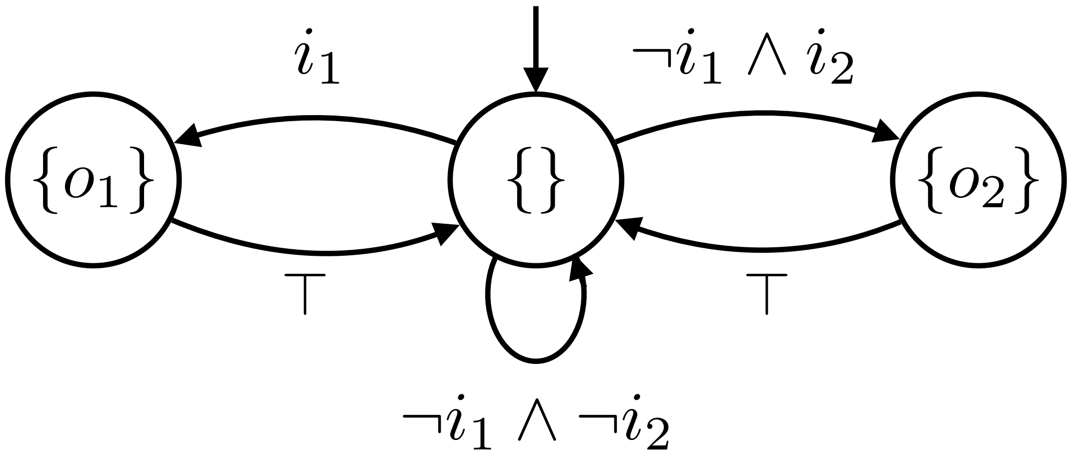
Counterfactuals alone are often imprecise.

Consider: $E = \bigcirc(o_1 \lor o_2)$ and $\pi = \{i_1, i_2\}\{o_1\}\{\}^\omega$.

$$\pi_{\neg i_1} = \{\cancel{i_1}, i_2\}\{o_2\}\{\}^\omega$$

$\implies$ $C = i_1$ alone does not negate the effect.

Contingencies on Traces



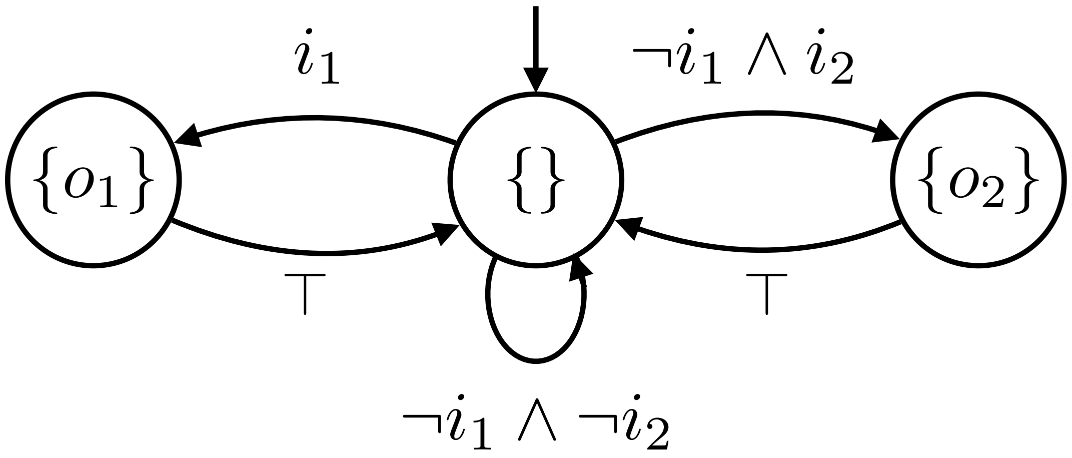
Counterfactuals alone are often imprecise.

Consider: $E = \bigcirc(o_1 \vee o_2)$ and $\pi = \{i_1, i_2\}\{o_1\}\{\}^\omega$.

$$\pi_{\neg(i_1 \vee i_2)} = \{\cancel{i_1}, \cancel{i_2}\}\{\}\{\}^\omega$$

$\implies C = i_1 \vee i_2$ works but is too imprecise.

Contingencies on Traces



Counterfactuals alone are often imprecise.

Consider: $E = \bigcirc(o_1 \vee o_2)$ and $\pi = \{i_1, i_2\}\{o_1\}\{\}^\omega$.

$$\pi^{cont.}_{\neg i_1} = \{\cancel{i_1}, i_2\}\{\cancel{o_2}\}\{\}^\omega$$

Contingency resets value.

$\Longrightarrow$ $C = i_1$ with the *contingency* $\bigcirc \neg o_2$ works.

C is a Cause for E iff...

SAT: π satisfies C and E .

CF: For every counterfactual input sequence σ , there exists a contingency trace π' such that $\sigma =_{inputs} \pi'$ and π' does not satisfy E .

MIN: There is no smaller cause candidate that satisfies the above.

Minimality

SAT and **CF** define a lot of potential causes.

$\pi_{\neg(i_1 \vee i_2)} = \{\cancel{i_1}, \cancel{i_2}\}\{\}\{\}^\omega \implies C = i_1 \vee i_2$ works but is too imprecise.

$\pi^{cont.}_{\neg i_1} = \{\cancel{i_1}, i_2\}\{\cancel{o_2}\}\{\}^\omega \implies C = i_1$ with the *contingency* $\bigcirc \neg o_2$ works.

Solution: prefer semantically minimal properties as causes, i.e., check:

$$i_1 \rightarrow (i_1 \vee i_2)$$

C is a Cause for E iff...

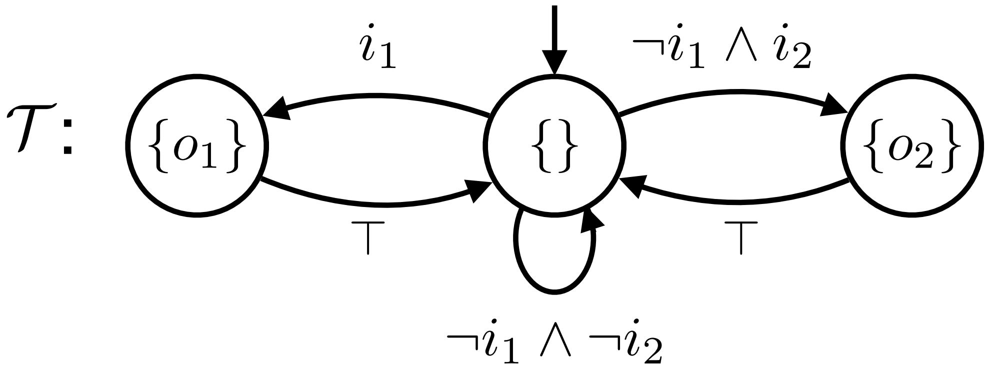
SAT: π satisfies C and E .

CF: For every counterfactual input sequence σ , there exists a contingency trace π' such that $\sigma =_{inputs} \pi'$ and π' does not satisfy E .

MIN: There does not exist a C' such that $C \rightarrow C'$ and C' satisfies **SAT** and **CF**.

Temporal Properties as Causes

$$\pi = (\{i_1, i_2\}\{i_1, i_2, o_1\})^\omega$$



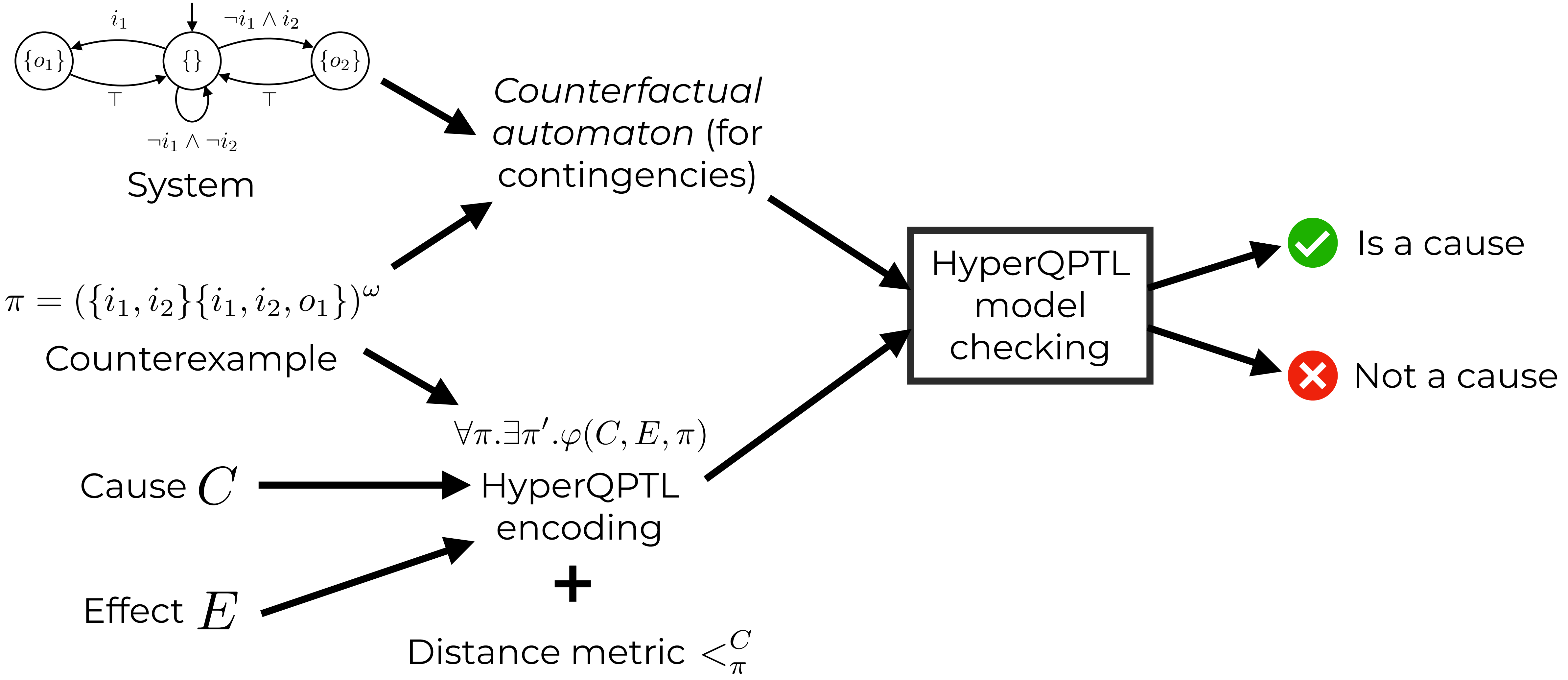
Does $\Diamond i_1$ cause $\Box \neg o_2$ on π in $\mathcal{T}$?

SAT: ✅

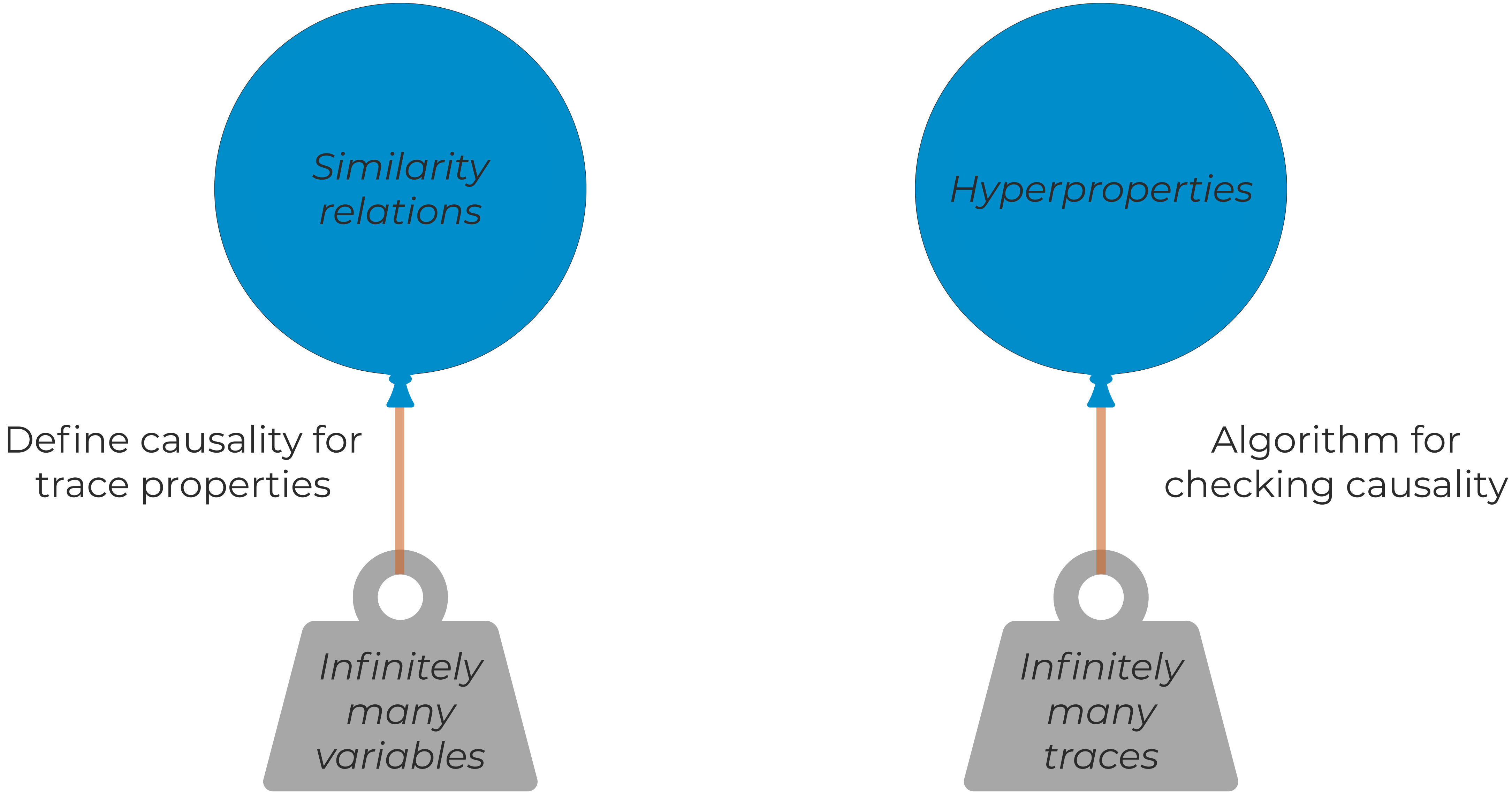
CF: ✅ $\pi' = (\{i_2\}\{i_2, o_2\})^\omega$ is the only counterfactual trace.

MIN: ❌ $\exists q.\ q \land \Box(\bigcirc q \leftrightarrow \neg q) \land \Box(q \rightarrow i_1)$ is more minimal (and the cause).

Checking Temporal Causality



Conclusion


Similarity relations
Define causality for trace properties
Infinitely many variables
Hyperproperties
Algorithm for checking causality
Infinitely many traces