Explaining Hyperproperty Violations[*]

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

[1] CISPA Helmholtz Center for Information Security
[2] Interactive Media Lab, Technische Universität Dresden

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. Fix-
ing the erroneous relations between traces in the system that led to the
counterexample is a difficult manual effort that highly benefits from addi-
tional 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 im-
proves on previous approaches for analyzing counterexamples returned
by HyperLTL model checkers.

1 Introduction

While model checking algorithms and tools (e.g., [19,18,25,13,54,46]) have, in
the past, focused on trace properties, recent failures in security-critical systems,
such as Heartbleed [27], Meltdown [58], Spectre [51], or Log4j [1], have triggered
the development of model checking algorithms for properties that relate multiple
computation traces to each other, i.e., *hyperproperties* [22]. Although the coun-
terexample returned by such a model checker for hyperproperties, which takes
the shape of a *set* of traces, may aid in the debugging process, understanding
and narrowing down which features are actually responsible for the erroneous
relation between the traces in the counterexample requires significantly more
manual effort than for trace properties. In this paper, we develop an explana-
tion technique for these more complex counterexamples that identifies the *actual
causes* [43,44,45] of hyperproperty violations.

Existing hyperproperty model checking approaches (e.g., [32,34,48]), take a
HyperLTL formula as an input. HyperLTL is a temporal logic extending LTL
with explicit trace quantification [21]. For example, observational determinism,

which requires that all traces π, π' agree on their observable outputs lo whenever they agree on their observable inputs li, can be formalized in HyperLTL as

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

In case a system model violates observational determinism, the model checker consequently returns a set of two execution traces witnessing the violation.

A first attempt in explaining model checking results of HyperLTL specifications has been made with HyperVis [47], which visualizes a counterexample returned by the model checker MCHyper [34] in a browser application. While the visualizations are already useful to analyze the counterexample at hand, it fails to identify causes for the violation in several security-critical scenarios. This is because HyperVis identifies important atomic proposition that appear in the HyperLTL formula and highlights these in the trace and the formula. For detecting causes, however, this is insufficient: a cause for a violation of observational determinism, for example, could be a branch on the valuation of a secret input i_s, which is not even part of the formula (see Sec. 3 for a running example).

Defining what constitutes an actual cause for an effect (a violation) in a given scenario is a precious contribution by Halpern and Pearl [43,44,45], who refined and formalized earlier approaches based on counterfactual reasoning [57]: Causes are sets of events such that, in the counterfactual world where they do not appear, the effect does not occur either. One of the main insights of Halpern and Pearl's work, however, is that naive counterfactuals are too imprecise. If, for instance, our actual cause preempted another potential cause, the mere absence of the actual cause will not be enough to prevent the effect, which will be still produced by the other cause in the new scenario. Halpern and Pearl's definition therefore allows to carefully control for other possible causes through the notion of *contingencies*. In the modified definition [43], contingencies allow to fix certain features of the counterfactual world to be exactly as they are in the actual world, regardless of the system at hand. Such a contingency effectively modifies the dynamics of the underlying model, and one insight of our work is that defining actual causality for reactive systems also needs to modify the system under a contingency. Notably, most works regarding trace causality [14,38] do not consider contingencies but only counterfactuals, and thus are not able to find true actual causes.

In this paper, we develop the notion of actual causality for effects described by HyperLTL formulas and use the generated causes as explanations for counterexamples returned by a model checker. We show that an implementation of our algorithm is practically feasible and significantly increases the state-of-the-art in explaining and analyzing HyperLTL model checking results. More precisely, our contributions and the outline of the paper are as follows:

- We define actual causality for violations of hyperproperties described by general HyperLTL formulas (Sec. 4).
- We present an efficient algorithm for computing actual causes for effects described in the practically relevant $\forall^*$-fragment of HyperLTL (Sec. 5).

– We report on the implementation of our algorithm and compare the generated explanations with the HyperVis visualisation tool (Sec. 6).

2 Preliminaries

We model a system as a *Moore machine* [61] $T = (S, s_0, AP, \delta, l)$ where S is a finite set of states, $s_0 \in S$ is the initial state, $AP = I \uplus O$ is the set of atomic propositions consisting of inputs I and outputs O, $\delta : S \times 2^I \to S$ is the transition function determining the successor state for a given state and input, and $l : S \to 2^O$ is the labeling function mapping each state to a set of outputs. A *trace* $t = t_0 t_1 t_2 \ldots \in (2^{AP})^\omega$ of T is an infinite sequence of sets of atomic propositions with $t_0 = l(s_0)$ and $t_i = A \cup l(s_i)$, where $A \subseteq I$ and $\delta(s_{i-1}, A) = s_i$ for all $i > 0$. With $traces(T)$, we denote the set of all traces of T. A trace property $P \subseteq T$ is a set of traces. A hyperproperty H is a lifting of a trace property, i.e., a *set of sets of traces*. A model satisfies a hyperproperty H if the set of traces of the model T is an element of the hyperproperty, i.e., $traces(T) \in H$.

2.1 HyperLTL

HyperLTL is a recently introduced logic for expressing temporal hyperproperties, extending linear-time temporal logic (LTL) [63] with trace quantification:

$$\varphi ::= \forall \pi. \varphi \mid \exists \pi. \varphi \mid \psi$$
$$\psi ::= a_\pi \mid \neg \psi \mid \psi \wedge \psi \mid \bigcirc \psi \mid \psi \,\mathcal{U}\, \psi \ .$$

We also consider the usual derived Boolean ($\vee$, $\to$, $\leftrightarrow$) and temporal operators ($\varphi \,\mathcal{R}\, \psi \equiv \neg(\neg\varphi \,\mathcal{U}\, \neg\psi)$, $\Diamond\varphi \equiv true\,\mathcal{U}\,\varphi$, $\Box\varphi \equiv false\,\mathcal{R}\,\varphi$). The semantics of HyperLTL formulas is defined with respect to a set of traces Tr and a trace assignment $\Pi : \mathcal{V} \to Tr$ that maps trace variables to traces. To update the trace assignment so that it maps trace variable π to trace t, we write $\Pi[\pi \mapsto t]$.

$$
\begin{array}{lll}
\Pi, i \vDash_{Tr} a_\pi & \text{iff} & a \in \Pi(\pi)[i] \\
\Pi, i \vDash_{Tr} \neg\varphi & \text{iff} & \Pi, i \nvDash_{Tr} \varphi \\
\Pi, i \vDash_{Tr} \varphi \wedge \psi & \text{iff} & \Pi, i \vDash_{Tr} \varphi \text{ and } \Pi, i \vDash_{Tr} \psi \\
\Pi, i \vDash_{Tr} \bigcirc\varphi & \text{iff} & \Pi, i+1 \vDash_{Tr} \varphi \\
\Pi, i \vDash_{Tr} \varphi \,\mathcal{U}\, \psi & \text{iff} & \exists j \geq i.\, \Pi, j \vDash_{Tr} \psi \wedge \forall i \leq k < j.\, \Pi, k \vDash_{Tr} \varphi \\
\Pi, i \vDash_{Tr} \exists \pi. \varphi & \text{iff} & \text{there is some } t \in Tr \text{ such that } \Pi[\pi \mapsto t], i \vDash_{Tr} \varphi \\
\Pi, i \vDash_{Tr} \forall \pi. \varphi & \text{iff} & \text{for all } t \in Tr \text{ it holds that } \Pi[\pi \mapsto t], i \vDash_{Tr} \varphi
\end{array}
$$

We explain counterexamples found by MCHYPER [34,24], which is a model checker for HyperLTL formulas, building on ABC [12]. MCHYPER takes as inputs a hardware circuit, specified in the AIGER format [8], and a HyperLTL formula. MCHYPER solves the model checking problem by computing the self-composition [6] of the system. If the system violates the HyperLTL formula, MCHYPER returns a counterexample. This counterexample is a set of traces through the original system that together violate the HyperLTL formula. Depending on the type of violation, this counterexample can then be used to debug the circuit or refine the specification iteratively.

2.2   Actual Causality

A formal definition of what actually causes an observed effect in a given context has been proposed by Halpern and Pearl [44]. Here, we outline the version later modified by Halpern [43]. Causality is defined with respect to a *causal model* $\mathcal{M} = (\mathcal{S}, \mathcal{F})$, given by a *signature* $\mathcal{S}$ and set of *structural equations* $\mathcal{F}$, which define the dynamics of the system. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ and $\mathcal{V}$ are disjoint sets of variables, termed *exogenous* and *endogenous* variables, respectively; and $\mathcal{R}$ defines the *range* of possible values $\mathcal{R}(Y)$ for all variables $Y \in \mathcal{U} \cup \mathcal{V}$. A *context* $\vec{u}$ is an assignment to the variables in $\mathcal{U} \cup \mathcal{V}$ such that the values of the exogenous variables are determined by factors outside of the model, while the value of some endogenous variable X is defined by the associated structural equation $f_X \in \mathcal{F}$. An *effect* φ in a causal model is a Boolean formula over assignments to endogenous variables. We say that a context $\vec{u}$ of a model $\mathcal{M}$ satisfies a partial variable assignment $\vec{X} = \vec{x}$ for $\vec{X} \subseteq \mathcal{U} \cup \mathcal{V}$ if the assignments in $\vec{u}$ and in $\vec{x}$ coincide for every variable $X \in \vec{X}$. The extension for Boolean formulas over variable assignments is as expected. For a context $\vec{u}$ and a partial variable assignment $\vec{X} = \vec{x}$, we denote by $(\mathcal{M}, \vec{u})[\vec{X} \leftarrow \vec{x}]$ the context $\vec{u}'$ in which the values of the variables in $\vec{X}$ are set according to $\vec{x}$, and all other values are computed according to the structural equations.

The actual causality framework of Halpern and Pearl aims at defining what events (given as variable assignments) are the cause for the occurrence if an effect in a specific given context. We now provide the formal definition.

Definition 1 ([44,43]). *A partial variable assignment $\vec{X} = \vec{x}$ is an* actual cause *of the effect φ in $(\mathcal{M}, \vec{u})$ if the following three conditions hold.*

AC1: $(\mathcal{M}, \vec{u}) \vDash \vec{X} = \vec{x}$ and $(\mathcal{M}, \vec{u}) \vDash \varphi$, i.e., both cause and effect are true in the actual world.

AC2: There is a set $\vec{W} \subseteq V$ of endogenous variables and an assignment $\vec{x}'$ to the variables in $\vec{X}$ such that if $(\mathcal{M}, \vec{u}) \vDash \vec{W} = \vec{w}$, then $(\mathcal{M}, \vec{u})[\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}] \vDash \neg\varphi$.

AC3: $\vec{X}$ is minimal, i.e. no subset of $\vec{X}$ satisfies AC1 and AC2.

Intuitively, AC2 states that in the counterfactual world obtained by intervening on the cause $\vec{X} = \vec{x}$ in the actual world (that is, setting the variables in $\vec{X}$ to $\vec{x}'$), the effect does not appear either. However, intervening on the possible cause might not be enough, for example when that cause preempted another. After intervention, this other cause may produce the effect again, therefore clouding the effect of the intervention. To address this problem, AC2 allows to reset certain parts of the counterfactual world back to their values in the actual world through the notion of *contingencies* (i.e., the set $\vec{W}$). For an illustration of Halpern and Pearl's actual causality, see Ex. 1 in Sec. 3.

3   Running Example

Consider a security-critical setting with two security levels: a high-security level h and a low-security level l. Inputs and outputs labeled as high-security, denoted

by hi and ho respectively, are confidential and thus only visible to the user itself, or, e.g., admins. Inputs and outputs labeled as low-security, denoted by li and lo respectively, are public and are considered to be observable by an attacker.

Our system of interest is modeled by the state graph representation shown in Fig. 1, which is treated as a black box by an attacker. The system is run without any low-security inputs, but branches depending on the given high-security inputs. If in one of the first two steps of an execution, a high-security input hi is encountered, the system outputs only the high-security variable ho directly afterwards and in the subsequent steps both outputs, regardless of inputs. If no high-security input is given in the first step, the low-security output lo is enabled and after the second step, again both outputs are enabled, regardless of what input is fed into the system.
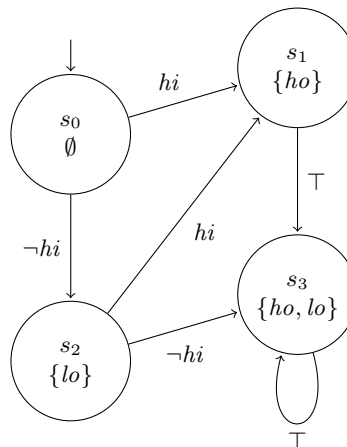
A prominent example hyperproperty is *observational determinism* from the introduction: which states that any sequence of low-inputs always produces the same low-outputs, regardless of what the high-security level inputs are. $\varphi = \forall\pi.\forall\pi'.\Box(li_\pi \leftrightarrow li_{\pi'}) \rightarrow$

Fig. 1: State graph representation of our example system.

$\Box(lo_\pi \leftrightarrow lo_{\pi'})$. The formula states that all traces π and π' must agree in the low-security outputs if they agree in the low-security inputs. Our system at hand does not satisfy observational determinism, because the low-security outputs in the first two steps depend on the present high-security inputs. Running MCHyper, a model checker for HyperLTL, results in the following counterexample:

$$t_1 = \{\}\{lo\}\{ho, lo\}^\omega \ ,$$
$$t_2 = \{hi\}\{hi, ho\}\{ho, lo\}^\omega \ .$$

With the same low-security input (none) the traces produce different low-security outputs by visiting either state s_1 or s_2 on the way to s_3.

In this paper, our goal is to explain the violation of a HyperlTL formula on such a counterexample. Following Halpern and Pearl's explanation framework [45], an actual cause that is considered to be possibly true or possibly false constitutes an explanation for the user. We only consider causes over input variables, which can be true and false in any model. Hence, finding an explanation amounts to answering which inputs caused the violation on a specific counterexample. Before we answer this question for HyperLTL and the corresponding counterexamples given by sets of traces (see Sec. 4), we first illustrate Halpern and Pearl's actual causality (see Sec. 2.2) with the above running example.

Example 1. Finite executions of a system can be modeled in Halpern and Pearl's causal models. We encode the first two execution steps of a single run of the

system. Consider inputs as exogenous variables $\mathcal{U} = \{hi_0, hi_1\}$ and outputs as endogenous variables $\mathcal{V} = \{lo_1, lo_2, ho_1, ho_2\}$. We have that $\mathcal{R}(Y) = \{0,1\}$ for every $Y \in \mathcal{U} \cup \mathcal{V}$. The structural equations encode the transitions: We have (1) $lo_1 = \neg hi_0$, (2) $ho_1 = hi_0$, (3) $lo_2 = \neg hi_1 \vee \neg lo_1$ and (4) $ho_2 = lo_1 \vee ho_1$. Consider context $\vec{u} = \{hi_0 = 0, hi_1 = 1\}$, effect $\varphi \equiv lo_1 = 1 \vee lo_2 = 1$, and candidate cause $hi_0 = 0$. Because of Eq. (1), we have that $(\mathcal{M}, \vec{u}) \vDash hi_0 = 0$ and $(\mathcal{M}, \vec{u}) \vDash lo_1 = 1$, hence AC1 is satisfied. Regarding AC2, this example allows us to illustrate the need for contingencies to accurately determine the actual cause: If we only consider intervening on the candidate cause $hi_0 = 0$, we still have $(\mathcal{M}, \vec{u})[hi_0 \leftarrow 1] \vDash \varphi$, because with $lo_1 = 0$ and Eq. (3) it follows that $(\mathcal{M}, \vec{u}) \vDash lo_2 = 1$. However, in the actual world, the second high input has no influence on the effect. We can control for this by considering the contingency $lo_2 = 0$, which is satisfied in the actual world, but not after the intervention on hi_0. Because of this contingency, we then have that $(\mathcal{M}, \vec{u})[hi_0 \leftarrow 0, lo_2 \leftarrow 0] \vDash \neg \varphi$, and hence, AC3 holds. This means we can infer that the first high input hi_0 was the actual cause for any low output to be enabled in the actual world. Note that, intuitively, the contingency allows us to ignore some of the structural equations by ignoring the value they assign to lo_2 in this context. Our definitions in Sec. 4 will allow similar modifications for counterexamples to hyperproperties.

4   Causality for Hyperproperty Violations

Our goal in this section is to formally define actual causality for the violation of a hyperproperty described by a general HyperLTL formula φ. Such a counterexample is given by a trace assignment to the trace variables appearing in φ. Note that, for universal quantifiers, the assignment of a single trace to the bounded variable suffices to define a counterexample. For existential quantifiers, this is not the case: to prove that an existential quantifier cannot be instantiated we need to show that no system trace satisfies the formula in its body, i.e., provide a proof for the whole system. In this work, we are interested in explaining violations of hyperproperties, and not proofs of their satisfaction [17]. Hence, we limit ourselves to instantiations of the outermost universal quantifiers of a HyperLTL formula, which can be returned by model checkers like MCHyper [34,24].

Definition 2 (Counterexample). *Let T be a transition system and denote $Traces(T) := Tr$, and let φ be a HyperLTL formula of the form $\forall \pi_1 \ldots \forall \pi_k \psi$, where ψ is a HyperLTL formula that does not start with $\forall$. A counterexample to φ in T is a partial trace assignment $\Gamma : \{\pi_1, \ldots, \pi_k\} \to Tr$ such that $\Gamma, 0 \vDash_{Tr} \neg \psi$.*

For ease of notation, we sometimes refer to Γ simply as the tuple of its instantiations $\Gamma = \langle \Gamma(\pi_1), \ldots, \Gamma(\pi_k) \rangle$. In terms of Halpern and Pearl's actual causality as outlined in Sec. 2.2, a counterexample describes the actual world at hand, which we want to explain. As a next step, we need to define an appropriate language to reason about possible causes and contingencies in our counterexample. We will use sets of *events*, i.e., values of atomic propositions at a specific position of a specific trace in the counterexample.

Definition 3 (Event). *An event is a tuple $e = \langle l_a, n, t \rangle$ such that $t \in (2^{AP})^\omega$ is a trace of a system T, $n \in \mathbb{N}$ is a point in time, and $l_a = a$ or $l_a = \neg a$ for an atomic proposition $a \in AP$. We say that a counterexample $\Gamma = \langle t_1, \ldots t_k \rangle$ satisfies a set of events $\mathcal{C}$, and denote $\Gamma \vDash \mathcal{C}$, if for every event $\langle l_a, n, t \rangle \in \mathcal{C}$ the two following conditions hold:*

1. *$t = t_i$ for some $i \in [1, k]$. That is, all events in $\mathcal{C}$ reason about traces in Γ*
2. *$l_a = a$ iff $a \in t[n]$, that is, iff a holds at time step n on t.*

We assume that the set AP is a disjoint union of input an output propositions, that is, $AP = I \uplus O$. We say that $\langle l_a, n, t \rangle$ is an *input event* if $a \in I$, and we call it an *output event* if $a \in O$. We denote the set of input events by IE and the set of output events by OE. These events have a direct correspondence with the variables appearing in Halpern and Pearl's causal models: we can identify input events with exogenous variables (because their value is determined by factors outside of the system) and output events with endogenous variables.

We define a cause as a set of input events, while an effect is a possibly infinite Boolean formula over OE. Note that, similar to [36], every HyperLTL formula can be represented as a first order formula over events,.e.g, $\forall \pi \forall \pi' \square (a_\pi \leftrightarrow a_{\pi'}) = \forall \pi \forall \pi' \bigwedge_{n \in \mathbb{N}} (\langle a, n, \pi \rangle \leftrightarrow \langle a, n, \pi' \rangle)$. For some set of events $\mathcal{S}$, let $^+\mathcal{S}_\pi^k = \{a \in AP \mid \langle a, k, \pi \rangle \in \mathcal{S}\}$ denote the set of atomic propositions defined positively by $\mathcal{S}$ on trace π at position k. Dualy, we define $^-\mathcal{S}_\pi^k = \{a \in AP \mid \langle \neg a, k, \pi \rangle \in \mathcal{S}\}$.

In order to define actual causality for hyperproperties we need to formally define how we obtain the counterfactual executions under some contingency for the case of events on infinite traces. We define a contingency as a set of output events. Mapping Halpern and Pearl's definition to transition systems, contingencies reset outputs in the counterfactual traces back to their value in the original counterexample, which amounts to changing the state of the system, and then following the transition function from the new state. For a given trace of the counterexample, we describe all possible behaviors under *arbitrary* contingencies with the help of a counterfactual automaton. The concrete contingency on a trace is defined by additional input variables. In the following, let $IC = \{ o^C \mid o \in O \}$ be a set of auxiliary input variables and $c : O \to IC$ be a function s.t. $c(o) = o^C$.

Definition 4 (Counterfactual Automaton). *Let $T = (S, s_0, AP, \delta, l)$ be a system with $S = 2^O$, i.e., every state is uniquely labeled, and there exists a state for every combination of outputs. Let $\pi = \pi_0 \ldots \pi_i (\pi_j \ldots \pi_n)^\omega \in traces(T)$ be a trace of T in a finite, lasso-shaped representation. The counterfactual automaton $T_\pi^C = (S \times \{0 \ldots n\}, (s_0, 0), (IC \cup I) \cup (O \cup \{0 \ldots n\}), \delta^C, l^C)$ is defined as follows:*

- *$\delta^C((s, k), Y) = (s', k')$ where $k' = j$ if $k = n$, else $k' = k + 1$, and*
  $l(s') = \{o \in O \mid (o \in \delta(s, Y \cap I) \wedge c(o) \notin Y) \vee (o \in \pi_{k'} \wedge c(o) \in Y)\}$,
- *$l^C(s, k) = l(s) \cup \{k\}$.*

A counterfactual automaton is effectively a chain of copies of the original system, of the same length as the counterexample. An execution through the counterfactual automaton starts in the first copy corresponding to the first position in the counterexample trace, and then moves through the chain until it

eventually loops back from copy n to copy j. A transition in the counterfactual automaton can additionally specify setting as a contingency some output variable o if the auxiliary input variable o^C is enabled. In this case, the execution will move to a state in the next automaton of the chain where all the outputs are as usual, except o, which will have the same value as in the counterexample π. Note that, under the assumption that all states of the original system are uniquely labeled and there exists a state for every combination of output variables, the function δ^C is uniquely determined. A counterfactual automaton for our running example is described in App. A.3.

Next, we need to define how we intervene on a set of traces with a candidate cause given as a set of input events, and a contingency given as a set of output events. We define an intervention function, which transforms a trace of our original automaton to an input sequence of an counterfactual automaton.

Definition 5 (Intervention). *For a cause $\mathcal{C} \subseteq IE$, a contingency $\mathcal{W} \subseteq OE$ and a trace π , we define the function intervene : $2^{AP} \times 2^{IE} \times 2^{OE} \to 2^{I \cup IC}$ as follows:*

$$intervene(\pi, \mathcal{C}, \mathcal{W})[k] = (\pi[k] \setminus {}^+\mathcal{C}_\pi^k) \cup {}^-\mathcal{C}_\pi^k \cup \{c(o) \mid o \in {}^+\mathcal{W}_\pi^k \cup {}^-\mathcal{W}_\pi^k\} \ .$$

We lift the intervention function to counterexamples given as a tuple $\Gamma = \langle \pi_1, \ldots, \pi_k \rangle$ as follows:

$$intervene(\Gamma, \mathcal{C}, \mathcal{W}) = \langle T_{\pi_1}^C(intervene(\pi_1, \mathcal{C}, \mathcal{W})), \ldots, T_{\pi_k}^C(intervene(\pi_k, \mathcal{C}, \mathcal{W})) \rangle \ .$$

Intuitively, the intervention function *flips* all the events that appear in the cause Γ: If some $a \in I$ appears positively in the candidate cause $\mathcal{C}$, it will appear negatively in the resulting input sequence, and vice-versa. For a contingency $\mathcal{W}$, the intervention function enables their auxiliary input for the counterfactual automaton at the appropriate time point irrespective of their value, as the counterfactual automaton will take care of matching the atomic propositions value to the value in the original counterexample Γ.

4.1   Actual Causality for HyperLTL Violations

We are now ready to formalize what constitutes an actual cause for the violation of a hyperproperty described by a HyperLTL formula.

Definition 6 (Actual Causality for HyperLTL). *Let Γ be a counterexample to a HyperLTL formula φ in a system T. The set $\mathcal{C}$ is an actual cause for the violation of φ on Γ if the following conditions hold.*

SAT $\Gamma \vDash \mathcal{C}$.
CF *There exist a contingency $\mathcal{W}$ and a non empty subset $\mathcal{C}' \subseteq \mathcal{C}$ such that:*

$$\Gamma \vDash \mathcal{W} \wedge intervene(\Gamma, \mathcal{C}', \mathcal{W}) \vDash_{traces(T)} \varphi \ .$$

MIN $\mathcal{C}$ *is minimal, i.e., no subset of $\mathcal{C}$ satisfies SAT and CF.*

Unlike in Halpern and Pearl's definition (see Sec. 2.2), the condition SAT requires Γ to satisfy only the cause, as we already know that the effect $\neg\varphi$, i.e., the violation of the specification, is satisfied by virtue of Γ being a counterexample. CF is the counterfactual condition corresponding to AC2 in Halpern and Pearl's definition, and it states that after intervening on the cause, under a certain contingency, the set of traces satisfies the property. MIN is the minimality criterion directly corresponding to AC3.

Example 2. Consider our running example from Sec. 3, i.e., the system from Fig. 1 and the counterexample to observational determinism $\Gamma = \langle t_1, t_2 \rangle$. Let us consider what it means to intervene on the cause $\mathcal{C}_1 = \{\langle hi, 0, t_2 \rangle\}$. Note that we have $\Gamma \vDash \mathcal{C}_1$, hence the condition SAT is satisfied. For CF, let us first consider an intervention without contingencies. This results in

$$intervene(\Gamma, \mathcal{C}_1, \emptyset) = \langle t_1', t_2' \rangle = \langle t_1, \{\}\{hi, lo\}\{ho\}\{ho, lo\}^\omega \rangle \ .$$

However, we still have $intervene(\Gamma, \mathcal{C}_1, \emptyset) \vDash_{traces(T)} \neg\varphi$, because the low outputs of t_1' and t_2' differ at the third position: $lo \in t_1'[2]$ and $lo \notin t_2'[2]$. This is because now the second high input takes effect, which was preempted by the first cause in the actual counterexample. The contingency $\mathcal{W}_2 = \{\langle lo, 2, t_2 \rangle\}$ now allows us to control for this by *modyfing the state* after taking the second high input as follows:

$$intervene(\Gamma, \mathcal{C}_2, \mathcal{W}_2)) = \langle t_1'', t_2'' \rangle = \langle t_1, \{\}\{hi, lo\}\{ho, lo\}\{ho, lo\}^\omega \rangle \ .$$

Note that t_2'' is not a trace of the model depicted in Fig. 1, because there is no transition that explains the step from $t_2''[1]$ to $t_2''[2]$. It is, however, a trace of the counterfactual automaton $T_{t_2}^C$ (see App. A.3), which encodes the set of counterfactual worlds for the trace t_2. The fact that we consider executions that are not part of the original system allows us to infer that only the first high input was an actual cause in our running example. Disregarding contingencies, we would need to consider both high inputs as an explanation for the violation of observational determinism, even though the second high input had no influence. Our treatment of contingencies corresponds directly to Halpern and Pearl's causal models, which allow to ignore certain structural equations as outlined in Ex. 1.

4.2   Finding Actual Causes with Model Checking

In this section, we consider the relationship between finding an actual cause for the violation of a HyperLTL formula starting with a universal quantifier and model checking of HyperLTL. We show that the problem of finding an actual cause can be reduced to a model checking problem where the generated formula for the model checking problem has one additional quantifier alternation. While there might be a reduction resulting in a more efficient encoding, our current result suggests that causality checking is the harder problem. The key idea of our reduction is to use counterfactual automata (that encode the given counterexample and the possible counterfactual traces) together with the HyperLTL

formula described in the proof to ensure the conditions SAT, CF, and MIN on the witnesses for the model checking result.

Proposition 1. *We can reduce the problem of finding an actual cause for the violation of a general HyperLTL formula starting with a universal quantifier to the model checking problem of a HyperLTL formula with one additional quantifier alternation.*

Proof. Let $\Gamma = \langle t_1, \ldots t_k \rangle$ be a counterexample for the formula $\forall \pi_1 \ldots \forall \pi_k. \varphi$ where φ is a HyperLTL formula that does not have a universal first quantifier. We provide the proof for the case of $\Gamma = \langle t_1, t_2 \rangle$ for readability reasons, but it can be extended to any natural number k. We assume that t_1, t_2 have some ω-regular representation, as otherwise the initial problem of computing causality is not well defined. That is, we denote $t_i = u_i(v_i)^\omega$ such that $|u_i \cdot v_i| = n_i$.

In order to find an actual cause, we need to find a pair of traces t_1', t_2' that are counterfactuals for t_1, t_2; satisfy the property φ; and the changes from t_1, t_2 to t_1', t_2' are minimal with respect to set containment. Changes in inputs between t_i and t_i' in the loop part v_i should reoccur in t_i' repeatedly. Note that the differences between the counterexample t_1, t_2 and the witnesses of the model checking problem t_1', t_2' encode the actual cause, i.e. in case of a difference, the cause contains the event that is present on the counterexample. To reason about these changes, we use the counterfactual automaton T_i^C for each t_i. Note that each T_i^C constitutes of n_i copies, that indicate in which step the automaton is, with respect to t_i and its loop v_i. For $m > |u_i|$, we label each state (s_i, m) in T_i^C with the additional label $L_{s_m,i}$, to indicate that the system is now in the loop part of t_i. In addition, we add the initial state of T_i^C the label l_i, and we add the initial state of the system T the label l_{or}. The formula ψ_{loop}^i below states that the trace π begins its run from the initial state of T_i^C (and thus stays in this component through the whole run), and that every time π visits a state on the loop, the same input sequence is observed. This way we enforce the periodic input behavior of the traces t_1, t_2 on t_1', t_2'.

$$\psi_{loop}^i(\pi) := l_{i,\pi} \wedge \bigwedge_{L_{s_m,i}} \bigvee_{A \subseteq I} \Box(L_{s_m,i,\pi} \to (\bigwedge_{a \in A} a_\pi \wedge \bigwedge_{a \notin A} \neg a_\pi))$$

For a subset of locations $N \subseteq [1, n_i]$ and a subset of input propositions $A \subseteq I$ we define $\psi_{diff}^i[N, A](\pi)$ that states that π differs from t_i in at least all events $\langle l_a, m, t_i \rangle$ for $a \in A, m \in N$; and the formula $\psi_{eq}^i[N, A](\pi)$ that states that for all events that are not defined by A and N, π is equal to t_i.

$$\psi_{diff}^i[N, A](\pi) = \bigwedge_{j \in N, a \in A} \bigcirc^j (a_\pi \not\leftrightarrow a_{t_i})$$

$$\psi_{eq}^i[N, A](\pi) = \bigwedge_{j \notin N, a \in I} \bigcirc^j (a_\pi \leftrightarrow a_{t_i}) \wedge \bigwedge_{j \in [1, n_i], a \notin A} \bigcirc^j (a_\pi \leftrightarrow a_{t_i})$$

We now define the formula ψ^i_{min} that states that the set of inputs (and locations) on which trace π differs from t_i is not contained in the corresponding set for π'. We only check locations up until the length n_i of t_i.

$$\psi^i_{min}(\pi, \pi') := \bigwedge_{N \subseteq [i, n_i]} \bigwedge_{A \subseteq \mathcal{I}} \left((\psi^i_{diff}[N, A](\pi) \wedge \psi^i_{eq}[N, A](\pi)) \to \neg\psi^i_{eq}[N, A](\pi') \right)$$

Denote $\varphi := Q_1\tau_1 \ldots Q_n\tau_n.\ \varphi'(\pi_1, \pi_2)$. The formula ψ_{cause} described below states that the two traces π'_1 and π'_2 are part of the systems T^C_1, T^C_2, and have the same loop structure as t_1 and t_2, and satisfy φ. That is, these traces can be obtained by changing the original traces t_1, t_2 and avoid the violation.

$$\psi_{cause}(\pi'_1, \pi'_2) := \varphi'(\pi'_1, \pi'_2) \wedge \bigwedge_{i=1,2} \psi^i_{loop}(\pi'_i)$$

Finally, ψ_{actual} described below states that the counterfactuals π'_1, π'_2 correspond to a minimal change in the input events with respect to t_1, t_2. All other traces that the formula reasons about start at the initial state of the original system and thus are not affected by the counterfactual changes. We verify ψ_{actual} against the product automaton $T \times T^C_1 \times T^C_2$ to find these traces $\pi'_i \in T^C_i$ that witness the presence of a cause, counterfactual and contingency.

$$\psi_{actual} := \exists\pi'_1\exists\pi'_2.\ \forall\pi''_1\pi''_2.\ Q_1\tau_1 \ldots Q_n\tau_n.\ \psi_{cause}(\pi'_1, \pi'_2) \wedge \bigwedge_{i=1,2} (l_{i,\pi'_i} \wedge l_{i,\pi''_i})$$

$$\wedge \bigwedge_{i \in [1,n]} l_{or,\tau_i} \wedge \left(\psi_{cause}(\pi''_1, \pi''_2) \to \left(\bigwedge_{i=1,2} \psi^i_{min}(\pi'_i, \pi''_i) \right) \right)$$

Then, if there exists two such traces π'_1, π'_2 in the system $T \times T^C_1 \times T^C_2$, they correspond to a minimal cause for the violation. Otherwise, there are no traces of the counterfactual automata that can be obtained from t_1, t_2 using counterfactual reasoning and satisfy the formula φ. $\square$

We have shown that we can use model checking HyperLTL to find an actual cause for the violation of a HyperLTL formula. The resulting model checking problem has an additional quantifier alternation which suggests that identifying actual causes is a harder problem. Therefore, we restrict ourselves to finding actual causes for violations of HyperLTL formulas from the universal fragment. This keeps the algorithms we present in the next section practical.

5   Computing Causes for Counterexamples

In this section, we describe our algorithm for finding actual causes of hyperproperty violations. Our algorithm is implemented on top of MCHyper [34], a model checker for hardware circuits and the alternation-free fragment of HyperLTL. In

case of a violation, our analysis enriches the provided counterexample with the actual cause which can explain the reason for the violaiton to the user.

We first provide an overview of our algorithm and then discuss each step in detail. First, we compute an over-approximation of the cause using a satisfiability analysis over the transitions taken in the counterexample. This analysis results in a set of candidate events $\tilde{\mathcal{C}}$. As we show in Prop. 2, every actual cause $\mathcal{C}$ for the violation is a subset of $\tilde{\mathcal{C}}$. In addition, in Prop. 3 we show that the set $\tilde{\mathcal{C}}$ satisfies conditions SAT and CF. To ensure MIN, we search for the smallest subset $\mathcal{C} \subseteq \tilde{\mathcal{C}}$ that satisfies SAT and CF. This set $\mathcal{C}$ is then our minimal and therefore actual cause. To check condition CF, we need to check the counterfactual of each candidate cause $\mathcal{C}$, and potentially also look for contingencies for $\mathcal{C}$. We separate our discussion as follows. We first discuss the calculation of the over approximation $\tilde{\mathcal{C}}$ (Sec. 5.1), then we present the ActualCause algorithm that identifies a minimal subset of $\tilde{\mathcal{C}}$ that is an actual cause (Sec. 5.2), and finally we discuss in detail the calculation of contingencies (Sec. 5.3).

In the following sections, we use a reduction of the universal fragment of HyperLTL to LTL, and the advantages of the linear translation of LTL to alternating automata, as we now briefly outline.

HyperLTL to LTL. Let φ be a $\forall^n$-HyperLTL formula and Γ be the counterexample. We construct an LTL formula φ' from φ as follows [30]: atomic propositions indexed with different trace variables are treated as different atomic propositions and trace quantifiers are eliminated. For example $\forall \pi, \pi'.a_\pi \wedge a_{\pi'}$ results in the LTL formula $a_\pi \wedge a_{\pi'}$. As for Γ, we use the same renaming in order to zip all traces into a single trace, for which we assume the finite representation $t'' = u'' \cdot (v'')^\omega$, which is also the structure of the model checker's output. The trace t'' is a violation of the formula φ', i.e., t'' satisfies $\neg\varphi'$. We denote $\bar{\varphi} := \neg\varphi'$. We can then assume, for implementation concerns, that the specification (and its violation) is an LTL formula, and the counterexample is a single trace. After our causal analysis, the translation back to a cause over hyperproperties is straightforward as we maintain all information about the different traces in the counterexample.

Finite Trace Model Checking Using Alternating Automata. In verifying condition CF (that is, in computing counterfactuals and contingencies), we need to apply finite trace model checking, as we want to check if the modified trace in hand still violates the specification φ, that is, satisfies $\bar{\varphi}$. To this end, we use the linear algorithm of [35], that exploits the linear translation of $\bar{\varphi}$ to an alternating automaton $\mathcal{A}_{\bar{\varphi}}$, and using backwards analysis checks the satisfaction of the formula. An alternating automaton [67] generalizes non-deterministic and universal automata, and its transition relation is a Boolean function over the states. The run of alternating automaton is then a *tree run* that captures the conjunctions in the formula. We use the algorithm of [35] as a black box (see App. A.1 for a formal definition of alternating automata). For the computation of contingencies we use an additional feature of the algorithm of [35] – the algorithm returns an accepting run tree $\mathcal{T}$ of $\mathcal{A}_{\bar{\varphi}}$ on t'', with annotations of nodes

that represent atomic subformulas of $\bar{\varphi}$ that take part in the satisfaction of $\bar{\varphi}$. We use this feature also in Sec. 5.1 when calculating the set of candidate causes.

5.1   Computing the Set of Candidate Causes

The events that might been a part of a the cause to the violation are in-fact all events that appear on the counterexample, or, equivalently, all events that appear in u'' and v''. Note that due to the finite representation, this is a finite set of events. Yet, not all events in this set can cause the violation. In order to remove events that could not have be a part of the cause, we perform an analysis of the transitions of the system taken during the execution of t''. With this analysis we detect which events appearing in the trace locally cause the respective transitions, and thus might be part of the global cause. Events that did not trigger a transition in this specific trace cannot be a part of the cause. Note that causing a transition and being an actual cause are two different notions - actual causality is defined over the behaviour of the system, not on individual traces. We denote the over-approximation of the cause as $\tilde{\mathcal{C}}$. Formally, we represent each transition as a Boolean function over inputs and states. Let δ_n denote the formula representing the transition of the system taken when reading $t''[n]$, and let $c_{a,n,i}$ be a Boolean variable that corresponds to the event $\langle a_{t_i}, n, t'' \rangle$.[3] Denote $\psi_n^t = \bigwedge_{a_{t_i} \in t''[n]} c_{a,n,i} \wedge \bigwedge_{a_{t_i} \notin t''[n]} \neg c_{a,n,i}$, that is, ψ_n^t expresses the exact set of events in $t''[n]$. In order to find events that might trigger the transition δ_n, we check for the *unsatisfiable core* of $\psi_n = (\neg \delta_n) \wedge \psi_n^t$. Intuitively, the unsatisfiable core of ψ_n is the set of events that force the system to take this specific transition. For every $c_{a,n,i}$ (or $\neg c_{a,n,i}$) in the unsatisfiable core that is also a part of ψ_n^t, we add $\langle a, n, t_i \rangle$ (or $\langle \neg a, n, t_i \rangle$) to $\tilde{\mathcal{C}}$.

We use unsatisfiable cores in order to find input events that are necessary in order to take a transition. However, this might not be enough. There are cases in which inputs that appear in formula $\bar{\varphi}$ are not detected using this method, as they are not essential in order to take a transition; however, they might be considered a part of the actual cause, as negating them can avoid the violation. Therefore, as a second step, we apply the algorithm of [35] on the annotated automaton $\mathcal{A}_{\bar{\varphi}}$ in order to find the specific events that effect the satisfaction of $\bar{\varphi}$, and we add these events to $\tilde{\mathcal{C}}$. Then, the unsatisfiable core approach provides us with inputs that affect the computation and might cause the violation even though they do not appear on the formula itself; while the alternating automaton allows us to find inputs that are not essential for the computation, but might still be a part of the cause as they appear on the formula.

Proposition 2. *The set $\tilde{\mathcal{C}}$ is indeed an over approximation of the cause for the violation. That is, every actual cause $\mathcal{C}$ for the violation is a subset of $\tilde{\mathcal{C}}$.*

Proof (sketch). Let $e = \langle l_a, n, t \rangle$ be an event such that e is not in the unsatisfiable core of ψ_n and does not directly affect the satisfaction of $\bar{\varphi}$ according to the

---

[3] That is, $\neg c_{a,n,i}$ corresponds to the event $\langle \neg a_{t_i}, n, t'' \rangle$. Recall that the atomic propositions on the zipped trace t'' are annotated with the original trace t_i from Γ.

Algorithm 1: ActualCause($\varphi, \Gamma, \tilde{\mathcal{C}}$)

Input: Hyperproperty φ, counterexample Γ violating φ, and a set of candidate causes $\tilde{\mathcal{C}}$ for which conditions SAT and CF hold.

Output: A set of input events $\mathcal{C}$ which is an actual cause for the violation.

1 for $i \in [1, \ldots, |\tilde{\mathcal{C}}| - 1]$ **do**
2 | **for** $\mathcal{C} \subset \tilde{\mathcal{C}}$ *with* $|\mathcal{C}| = i$ **do**
3 | | let $\Gamma^f = intervene(\Gamma, \mathcal{C}, \emptyset)$;
4 | | **if** $\Gamma^f \vDash \varphi$ **then**
5 | | | **return** $\mathcal{C}$;
6 | | **else**
7 | | | $\tilde{\mathcal{W}} = $ ComputeContingency($\varphi, \mathcal{C}, \Gamma$);
8 | | | **if** $\tilde{\mathcal{W}} \neq \emptyset$ **then**
9 | | | | **return** $\mathcal{C}$;
10 return $\tilde{\mathcal{C}}$;

---

alternating automata analysis. That is, the transition corresponds to ψ_n^t is taken regardless of e, and thus all future events on t remain the same regardless of the valuation of e. In addition, the valuation of the formula $\bar{\varphi}$ is the same regardless of e, since: (1) e does not directly affect the satisfaction of $\bar{\varphi}$; (2) e does not affect future events on t (and obviously it does not affect past events). Therefore, every set $\mathcal{C}'$ such that $e \in \mathcal{C}'$ is not minimal, and does not form a cause. Since the above is true for all events $e \notin \mathcal{C}$, it holds that $\mathcal{C} \subseteq \tilde{\mathcal{C}}$ for every actual cause $\mathcal{C}$.   □

Proposition 3. *The set $\tilde{\mathcal{C}}$ satisfies conditions SAT and CF.*

Proof. The condition SAT is satisfied as we add to $\tilde{\mathcal{C}}$ only events that indeed occur on the counterexample trace. For CF, consider that $\tilde{\mathcal{C}}$ is a super-set of the actual cause $\mathcal{C}$, so the same contingency and counterfactual of $\mathcal{C}$ will also apply for $\tilde{\mathcal{C}}$. This is since in order to compute counterfactual we are allowed to flip any subset of the events in $\mathcal{C}$, and any such subset is also a subset of $\tilde{\mathcal{C}}$. In addition, in computing contingencies, we are allowed to flip any subset of outputs as long as they agree with the counterexample trace, which is independent in $\tilde{\mathcal{C}}$ and $\mathcal{C}$.   □

5.2   Checking Actual Causality

Due to Prop. 2 we know that in order to find an actual cause, we only need to consider subsets of $\tilde{\mathcal{C}}$ as candidate causes. In addition, since $\tilde{\mathcal{C}}$ satisfies condition SAT, so do all of its subsets. We thus only need to check conditions CF and MIN for subsets of $\tilde{\mathcal{C}}$. Our actual causality computation, presented in Alg. 1 is as follows. We start with the set $\tilde{\mathcal{C}}$, that satisfies SAT and CF. We then check if there exists a more minimal cause that satisfies CF. This is done by iterating over all subsets $\mathcal{C}'$ of $\tilde{\mathcal{C}}$, ordered by size and starting with the smallest ones, and checking if the counterfactual for the $\mathcal{C}'$ manages to avoid the violation; and if not, if there exists a contingency for this $\mathcal{C}'$. If the answer to one of these questions is yes, then $\mathcal{C}'$ is a minimal cause that satisfies SAT, CF, and MIN, and thus we return $\mathcal{C}'$ as our actual cause. We now elaborate on CF and MIN.

Algorithm 2: ComputeContingency($\varphi, \Gamma, \mathcal{C}$)

    Input: Hyperproperty φ, a counterexample Γ and a potential cause $\mathcal{C}$.
    Output: a set of output events $\mathcal{W}$ which is a contingency for φ, Γ and $\mathcal{C}$, or $\emptyset$
           if no contingency found.

1   **let** t'' be the zipped trace of Γ, φ' be the LTL formula obtained from φ, and
      $\bar{\varphi} = \neg\varphi'$;
2   **let** $\mathcal{A}_{\bar{\varphi}}$ be the alternating automaton for $\bar{\varphi}$;
3   **let** $\mathcal{N}$ be the sets of events derived from the annotated run tree of $\mathcal{A}_{\bar{\varphi}}$ on t^f;
4   **let** t^f be the counterfactual trace obtained from t'' by flipping all events in $\mathcal{C}$;
5   **let** $\mathcal{O}' := \{\langle l_{a_t}, n, t'' \rangle \in OE \mid a_t \in t''[n] \leftrightarrow a_t \notin t^f[n]\}$;
6   **for** *every subset* $\mathcal{W}' \subseteq (\mathcal{N} \cap \mathcal{O}')$, *and then for every other subset* $\mathcal{W}' \subseteq \mathcal{O}'$ **do**
7 | $t^m := intervene(t'', \mathcal{C}, \mathcal{W}')$;
8 | **if** $t^m \vDash \varphi'$ **then**
9 | | **return** $\mathcal{W}'$;
10 **return** $\emptyset$;

CF. As we have mentioned above, checking condition CF is done in two stages – checking for counterfactuals and computing contingencies. We first show that we do not need to consider all possible counterfactuals, but only one counterfactual for each candidate cause.

Proposition 4. *Let $\tilde{\mathcal{C}}$ be a candidate cause. Then, in order to check if $\tilde{\mathcal{C}}$ is an actual cause it is enough to test the one counterfactual where all the events in $\tilde{\mathcal{C}}$ are flipped.*

Proof. Assume that there is a strict subset $\mathcal{C}$ of $\tilde{\mathcal{C}}$ such that we only need to flip the valuations of events in $\mathcal{C}$ in order to find a counterfactual or contingency, thus $\mathcal{C}$ satisfies CF. Since $\mathcal{C}$ is a more minimal cause than $\tilde{\mathcal{C}}$, we will find it during the minimality check. □

    We assume that CF holds for the input set $\tilde{\mathcal{C}}$ and check if it holds for any smaller subset $\mathcal{C} \subset \tilde{\mathcal{C}}$. CF holds for $\mathcal{C}$ if (1) flipping all events in $\mathcal{C}$ is enough to avoid the violation of φ or if (2) there exists a non-empty set of contingencies for $\mathcal{C}$ that ensures that φ is not violated. The computation of contingencies is described in Alg. 2.

    Verifying condition CF involves model checking traces against an LTL formula, as we check if the property φ is still violated on the counterfactual trace with the empty contingency (line 3), and on the counterfactual traces resulting from the different contingency sets we consider in Alg. 2 (line 7). In both scenarios, we apply finite trace model checking, as described at the beginning of Sec. 5 (as we assume a finite, lasso-shaped representation of our traces).

MIN. To check if $\tilde{\mathcal{C}}$ is minimal, we need to check if there exists a subset of $\tilde{\mathcal{C}}$ that satisfies CF. We check CF for all subsets, starting with the smallest one, and report the first subset that satisfies CF as our actual cause. (Note that we already established that $\tilde{\mathcal{C}}$ and all of its subsets satisfy SAT.)

5.3   Computing Contingencies

Recall that the role of contingencies is to eliminate the effect of other possible causes from the counterfactual world, in case these causes did not affect the violation in the actual world. More formally, in computing contingencies we look for a set $\mathcal{W}$ of output events such that changing these outputs from their value in the counterfactual to their value in the counterexample t'' results in avoiding the violation. Note that the inputs remain as they are in the counterfactual. We note that the problem of finding contingencies is hard, and in general is equivalent to the problem of model checking. This is since we need to consider all traces that are the result of changing some subset of events (output + time step) from the counterfactual back to the counterexample, and to check if there exists a trace in this set that avoids the violation. Unfortunately, we are unable to avoid an exponential complexity in the size of the original system, in the worst case. However, our experiments show that in practice, most cases do not require the use of contingencies.

Our algorithm for computing contingencies (Alg. 2) works as follows. Let t^f be the counterfactual trace. As a first step, we use the annotated run tree $\mathcal{T}$ of the alternating automaton $\mathcal{A}_{\bar{\varphi}}$ on t^f to detect output events that appear in $\bar{\varphi}$ and take part in satisfying $\bar{\varphi}$. Subsets of these output events are our first candidates for contingencies as they directly related to the violation (Alg. 2 lines 3-9). If we were not able to find a contingency, we continue to checking all possible subsets of outputs events that differ from the original counterexample trace. We test the different outputs by feeding the Counterfactual automaton of Def. 4 with additional inputs from the set I^C. The resulted trace is then our candidate contingency, which we try to verify against φ. Note that we can bound the number of different input sequences by the size of the product of the Counterfactual automaton and the automaton for $\bar{\varphi}$, and thus the process terminates.

Theorem 1 (Correctness). *Our algorithm is sound and complete. That is, let Γ be a counterexample with a finite representation to a $\forall^n$-HyperLTL formula ψ. Then, our algorithm returns an actual cause for the violation, if such exists.*

Proof. Soundness. Since we verify each candidate set of inputs according to the conditions SAT, CF and MIN, it holds that every output of our algorithm is indeed an actual cause. *Completeness.* If there exists a cause, then due to Prop. 2, it is a subset of the finite set $\tilde{\mathcal{C}}$. Since in the worst case we test every subset of $\tilde{\mathcal{C}}$, if there exists a cause we will eventually find it.   □

6   Implementation and Experiments

We implemented Alg. 1 and evaluated it on publicly available example instances of HyperVis [47], for which their state graphs were available. In the following, we provide implementation details, report on the running time and show the usefulness of the implementation by comparing to the highlighting output of HyperVis.

Table 1: Experimental results of our implementation. Times are given in ms.

Instance	$\|\Gamma\|$	$\|\varphi\|$	time($\tilde{\mathcal{C}}$)	$\tilde{\mathcal{C}}$	#($\mathcal{C}$)	time($\forall\mathcal{C}$)
Running example	10	9	19	$\neg hi^0_{t_1}, hi^0_{t_2}$	2	55
Security in & out	35	19	292	$hi^2_{t_1}, \neg hi^0_{t_1}, \neg hi^3_{t_1}, \neg hi^1_{t_1}$ $hi^2_{t_2}, hi^0_{t_2}, hi^1_{t_2}, hi^3_{t_2}$	8	798
Drone example 1	24	19	33	$bound^2_{t_1}, \neg bound^1_{t_1}, up^1_{t_1}, \neg up^2_{t_1}$ $bound^2_{t_2}, \neg bound^1_{t_2}, \neg up^1_{t_2}$	5	367
Drone example 2	36	36	31	$bound^1_{t_1}, \neg bound^1_{t_2}, up^1_{t_2}$	3	256
Asymmetric arbiter '19	28	35	53	see Appendix A.4	10	490
Asymmetric arbiter	72	35	70	see Appendix A.4	24	1480

Our implementation is written in Python and uses py-aiger [59] and spot [26]. We compute the candidate cause according to Sec. 5.1 with py-sat [49], using Glucose 4 [3,65], building on Minisat [65]. We ran experiments on a MacBook Pro with a $3,3$ GHz Dual-Core Intel Core i7 processor and 16 GB RAM[4].

Experimental results. The results of our experimental valuation can be found in Tab. 1. We report on the size of the analyzed counterexample $\|\Gamma\|$, the size of the violated formula $\|\varphi\|$, how long it took to compute the first, over-approximated cause (see time($\tilde{\mathcal{C}}$)) and state the approximation $\tilde{\mathcal{C}}$ itself, the number of computed minimal causes #($\mathcal{C}$) and the time it took to compute all of them (see time($\forall\mathcal{C}$)). The `Running Example` is described in Sec. 3, the instance `Security in & out` refers to a system which leaks high security input by not satisfying a noninterference property, the `Drone examples` consider a leader-follower drone scenario, and the `Asymmetric Arbiter` instances refer to arbiter implementations that do not satisfy a symmetry constraint. Specifications can be found in App. A.5.

Our first observation is that the cause candidate $\tilde{\mathcal{C}}$ can be efficiently computed thanks to the iterative computation of unsatisfiable cores (Sec. 5.1). The cause candidate provides a tight over-approximation of possible minimal causes. As expected, the runtime for finding minimal causes increases for larger counterexamples. However, as our experiments show, the overhead is manageable, because we optimize the search for all minimal causes by only considering every subset in $\tilde{\mathcal{C}}$ instead of naively going over every combination of input events (see Prop. 2). Compared to the computationally heavy task of model checking to get a counterexample, our approach incurs little additional cost, which matches our theoretical results (see Prop. 1).

During our experiments, we have found that computing the candidate $\tilde{\mathcal{C}}$ first has, additionally to providing a powerful heuristic, another benefit: Even when the computation of minimal causes becomes increasingly expensive, $\tilde{\mathcal{C}}$ can serve as an intermediate result for the user. By filtering for important inputs, such as high security inputs, $\tilde{\mathcal{C}}$ already gives great insight to why the property was violated. In the asymmetric arbiter instance, for example, the input events

---

[4] After double-blind reviewing, code and results will be available online.

$\langle \neg tb\_secret, 3, t_0 \rangle$ and $\langle tb\_secret, 3, t_1 \rangle$ of $\tilde{\mathcal{C}}$, which cause the violation, immediately catch the eye (c.f App. A.4).

Comparison to HyperVis. HyperVis [47] is a tool for visualizing counterexamples returned from the HyperLTL model checker MCHyper [34]. It highlights the events in the trace that it considers responsible for the violation based on the formula and the set of traces, without considering the system model. However, violations of many relevant security policies such as observational determinism are not caused by events whose atomic propositions appear in the formula, as can be seen in our running example (see Sec. 3 and Ex. 2). When running the highlight function of HyperVis for the counterexample traces t_1, t_2 on `Running example`, the output events $\langle lo, 1, t_1 \rangle$ and $\langle \neg lo, 1, t_2 \rangle$ will be highlighted, neglecting the decisive high security input hi. Using our method additionally reveals the input events $\langle \neg hi, 0, t_1 \rangle$ and $\langle hi, 0, t_2 \rangle$, i.e., an actual cause (see Tab. 1). This pattern can be observed throughout all considered instances in our experiments. For instance in the `Asymmetric arbiter` instance mentioned above, the input events causing the violation do also not occur in the formula (see App. A.5) and thus HyperVis does not highlight this important cause for the violation.

7   Related Work

With the introduction of HyperLTL and HyperCTL* [21], temporal hyperproperties have been studied extensively: satisfiability [28,37,60], model checking [34,48,33], program repair [11], monitoring [31,10,66,2], synthesis [29], and expressiveness studies [23,52,36]. Causal analysis of hyperproperties has been studied theoretically based on counterfactual builders [39] instead of actual causality, as in our work. Explanation methods [4] exist for trace properties [68,5,38,40,41], integrated in several model checkers [20,15,16]. Minimization [53] has been studied, as well as analyzing several system traces together [42,64,9]. There exists work in explaining counterexamples for function block diagrams [50,62]. MOD-CHK uses a causality analysis [7] returning an over-approximation, while we provide minimal causes. Lastly, there are approaches which define actual causes for the violation of a trace property using Event Order Logic [14,55,56].

8   Conclusion

We present an explanation method for counterexamples to hyperproperties described by HyperLTL formulas. We lift Halpern and Pearl's definition of actual causality to effects described by hyperproperties and counterexamples given as sets of traces. Like the definition that inspired us, we allow modifications of the system dynamics in the counterfactual world through contingencies, and define these possible counterfactual behaviors in an automata-theoretic approach. The evaluation of our prototype implementation shows that our method is practically applicable and significantly improves the state-of-the-art in explaining counterexamples returned by a HyperLTL model checker.

References

1. Apache log4j security vulnerabilities, `https://logging.apache.org/log4j/2.x/security.html`
2. Agrawal, S., Bonakdarpour, B.: Runtime verification of k-safety hyperproperties in hyperltl. In: IEEE 29th Computer Security Foundations Symposium, CSF 2016, Lisbon, Portugal, June 27 - July 1, 2016. pp. 239–252. IEEE Computer Society (2016). https://doi.org/10.1109/CSF.2016.24, `https://doi.org/10.1109/CSF.2016.24`
3. Audemard, G., Simon, L.: Predicting learnt clauses quality in modern SAT solvers. In: Boutilier, C. (ed.) IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009. pp. 399–404 (2009), `http://ijcai.org/Proceedings/09/Papers/074.pdf`
4. Baier, C., Dubslaff, C., Funke, F., Jantsch, S., Majumdar, R., Piribauer, J., Ziemek, R.: From Verification to Causality-Based Explications. In: Bansal, N., Merelli, E., Worrell, J. (eds.) 48th International Colloquium on Automata, Languages, and Programming (ICALP 2021). Leibniz International Proceedings in Informatics (LIPIcs), vol. 198, pp. 1:1–1:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2021). https://doi.org/10.4230/LIPIcs.ICALP.2021.1, `https://drops.dagstuhl.de/opus/volltexte/2021/14070`
5. Ball, T., Naik, M., Rajamani, S.K.: From symptom to cause: localizing errors in counterexample traces. In: Aiken, A., Morrisett, G. (eds.) Conference Record of POPL 2003: The 30th SIGPLAN-SIGACT Symposium on Principles of Programming Languages, New Orleans, Louisisana, USA, January 15-17, 2003. pp. 97–105. ACM (2003). https://doi.org/10.1145/604131.604140, `https://doi.org/10.1145/604131.604140`
6. Barthe, G., D'Argenio, P.R., Rezk, T.: Secure information flow by self-composition. Math. Struct. Comput. Sci. **21**(6), 1207–1252 (2011). https://doi.org/10.1017/S0960129511000193, `https://doi.org/10.1017/S0960129511000193`
7. Beer, I., Ben-David, S., Chockler, H., Orni, A., Trefler, R.J.: Explaining counterexamples using causality. In: Bouajjani, A., Maler, O. (eds.) Computer Aided Verification, 21st International Conference, CAV 2009, Grenoble, France, June 26 - July 2, 2009. Proceedings. Lecture Notes in Computer Science, vol. 5643, pp. 94–108. Springer (2009). https://doi.org/10.1007/978-3-642-02658-4_11, `https://doi.org/10.1007/978-3-642-02658-4_11`
8. Biere, A.: The AIGER And-Inverter Graph (AIG) format version 20071012. Tech. Rep. Report 07/1, Institute for Formal Models and Verification, Johannes Kepler University, Altenbergerstr. 69, 4040 Linz, Austria (2007)
9. Bochot, T., Virelizier, P., Waeselynck, H., Wiels, V.: Paths to property violation: A structural approach for analyzing counter-examples. In: 12th IEEE High Assurance Systems Engineering Symposium, HASE 2010, San Jose, CA, USA, November 3-4, 2010. pp. 74–83. IEEE Computer Society (2010). https://doi.org/10.1109/HASE.2010.15, `https://doi.org/10.1109/HASE.2010.15`
10. Bonakdarpour, B., Finkbeiner, B.: The complexity of monitoring hyperproperties. In: 31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018. pp. 162–174 (2018). https://doi.org/10.1109/CSF.2018.00019, `https://doi.org/10.1109/CSF.2018.00019`

11. Bonakdarpour, B., Finkbeiner, B.: Program repair for hyperproperties. In: Automated Technology for Verification and Analysis - 17th International Symposium, ATVA 2019, Taipei, Taiwan, October 28-31, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11781, pp. 423–441. Springer (2019). https://doi.org/10.1007/978-3-030-31784-3_25, https://doi.org/10.1007/978-3-030-31784-3_25

12. Brayton, R., Mishchenko, A.: Abc: An academic industrial-strength verification tool. In: Touili, T., Cook, B., Jackson, P. (eds.) Computer Aided Verification. pp. 24–40. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)

13. Brayton, R.K., Mishchenko, A.: ABC: an academic industrial-strength verification tool. In: Touili, T., Cook, B., Jackson, P.B. (eds.) Computer Aided Verification, 22nd International Conference, CAV 2010, Edinburgh, UK, July 15-19, 2010. Proceedings. Lecture Notes in Computer Science, vol. 6174, pp. 24–40. Springer (2010), https://doi.org/10.1007/978-3-642-14295-6_5

14. Caltais, G., Guetlein, S.L., Leue, S.: Causality for general ltl-definable properties. In: Finkbeiner, B., Kleinberg, S. (eds.) Proceedings 3rd Workshop on formal reasoning about Causation, Responsibility, and Explanations in Science and Technology, CREST@ETAPS 2018, Thessaloniki, Greece, 21st April 2018. EPTCS, vol. 286, pp. 1–15 (2018). https://doi.org/10.4204/EPTCS.286.1, https://doi.org/10.4204/EPTCS.286.1

15. Chaki, S., Clarke, E.M., Groce, A., Jha, S., Veith, H.: Modular verification of software components in C. IEEE Trans. Software Eng. $30(6)$, 388–402 (2004). https://doi.org/10.1109/TSE.2004.22, https://doi.org/10.1109/TSE.2004.22

16. Chaki, S., Groce, A., Strichman, O.: Explaining abstract counterexamples. In: Taylor, R.N., Dwyer, M.B. (eds.) Proceedings of the 12th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2004, Newport Beach, CA, USA, October 31 - November 6, 2004. pp. 73–82. ACM (2004). https://doi.org/10.1145/1029894.1029908, https://doi.org/10.1145/1029894.1029908

17. Chockler, H., Halpern, J.Y., Kupferman, O.: What causes a system to satisfy a specification? ACM Trans. Comput. Log. $9(3)$, 20:1–20:26 (2008). https://doi.org/10.1145/1352582.1352588, https://doi.org/10.1145/1352582.1352588

18. Clarke, E.M., Biere, A., Raimi, R., Zhu, Y.: Bounded model checking using satisfiability solving. Formal Methods Syst. Des. $19(1)$, 7–34 (2001). https://doi.org/10.1023/A:1011276507260, https://doi.org/10.1023/A:1011276507260

19. Clarke, E.M., Emerson, E.A.: Design and synthesis of synchronization skeletons using branching-time temporal logic. In: Kozen, D. (ed.) Logics of Programs, Workshop, Yorktown Heights, New York, USA, May 1981. Lecture Notes in Computer Science, vol. 131, pp. 52–71. Springer (1981). https://doi.org/10.1007/BFb0025774, https://doi.org/10.1007/BFb0025774

20. Clarke, E.M., Kroening, D., Lerda, F.: A tool for checking ANSI-C programs. In: Jensen, K., Podelski, A. (eds.) Tools and Algorithms for the Construction and Analysis of Systems, 10th International Conference, TACAS 2004, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2004, Barcelona, Spain, March 29 - April 2, 2004, Proceedings. Lecture Notes in Computer Science, vol. 2988, pp. 168–176. Springer (2004). https://doi.org/10.1007/978-3-540-24730-2_15, https://doi.org/10.1007/978-3-540-24730-2_15

21. Clarkson, M.R., Finkbeiner, B., Koleini, M., Micinski, K.K., Rabe, M.N., Sánchez, C.: Temporal logics for hyperproperties. In: Abadi, M., Kremer, S. (eds.) Principles of Security and Trust - Third International Conference, POST 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, April 5-13, 2014, Proceedings. Lecture Notes in Computer Science, vol. 8414, pp. 265–284. Springer (2014). https://doi.org/10.1007/978-3-642-54792-8_15, `https://doi.org/10.1007/978-3-642-54792-8_15`

22. Clarkson, M.R., Schneider, F.B.: Hyperproperties. J. Comput. Secur. **18**(6), 1157–1210 (2010). https://doi.org/10.3233/JCS-2009-0393, `https://doi.org/10.3233/JCS-2009-0393`

23. Coenen, N., Finkbeiner, B., Hahn, C., Hofmann, J.: The hierarchy of hyperlogics. In: 34th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2019, Vancouver, BC, Canada, June 24-27, 2019. pp. 1–13. IEEE (2019). https://doi.org/10.1109/LICS.2019.8785713, `https://doi.org/10.1109/LICS.2019.8785713`

24. Coenen, N., Finkbeiner, B., Sánchez, C., Tentrup, L.: Verifying hyperliveness. In: Dillig, I., Tasiran, S. (eds.) Computer Aided Verification - 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11561, pp. 121–139. Springer (2019). https://doi.org/10.1007/978-3-030-25540-4_7, `https://doi.org/10.1007/978-3-030-25540-4_7`

25. Dehnert, C., Junges, S., Katoen, J., Volk, M.: A storm is coming: A modern probabilistic model checker. In: Majumdar, R., Kuncak, V. (eds.) Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II. Lecture Notes in Computer Science, vol. 10427, pp. 592–600. Springer (2017), `https://doi.org/10.1007/978-3-319-63390-9_31`

26. Duret-Lutz, A., Lewkowicz, A., Fauchille, A., Michaud, T., Renault, E., Xu, L.: Spot 2.0 - A framework for LTL and \omega -automata manipulation. In: Artho, C., Legay, A., Peled, D. (eds.) Automated Technology for Verification and Analysis - 14th International Symposium, ATVA 2016, Chiba, Japan, October 17-20, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9938, pp. 122–129 (2016), `https://doi.org/10.1007/978-3-319-46520-3_8`

27. Durumeric, Z., Li, F., Kasten, J., Amann, J., Beekman, J., Payer, M., Weaver, N., Adrian, D., Paxson, V., Bailey, M., et al.: The matter of heartbleed. In: Proceedings of the 2014 conference on internet measurement conference. pp. 475–488 (2014)

28. Finkbeiner, B., Hahn, C.: Deciding hyperproperties. In: 27th International Conference on Concurrency Theory, CONCUR 2016, August 23-26, 2016, Québec City, Canada. LIPIcs, vol. 59, pp. 13:1–13:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2016). https://doi.org/10.4230/LIPIcs.CONCUR.2016.13, `https://doi.org/10.4230/LIPIcs.CONCUR.2016.13`

29. Finkbeiner, B., Hahn, C., Lukert, P., Stenger, M., Tentrup, L.: Synthesis from hyperproperties. Acta Informatica **57**(1-2), 137–163 (2020). https://doi.org/10.1007/s00236-019-00358-2, `https://doi.org/10.1007/s00236-019-00358-2`

30. Finkbeiner, B., Hahn, C., Stenger, M., Tentrup, L.: RVHyper: A runtime verification tool for temporal hyperproperties. In: Proceedings of TACAS. LNCS, vol. 10806, pp. 194–200. Springer (2018). https://doi.org/10.1007/978-3-319-89963-3_11

31. Finkbeiner, B., Hahn, C., Stenger, M., Tentrup, L.: Monitoring hyperproperties. Formal Methods Syst. Des. **54**(3), 336–363 (2019).

https://doi.org/10.1007/s10703-019-00334-z,   `https://doi.org/10.1007/s10703-019-00334-z`

32. Finkbeiner, B., Hahn, C., Torfah, H.: Model checking quantitative hyperproperties. In: Computer Aided Verification - 30th International Conference, CAV 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17, 2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10981, pp. 144–163. Springer (2018). https://doi.org/10.1007/978-3-319-96145-3_8, `https://doi.org/10.1007/978-3-319-96145-3_8`

33. Finkbeiner, B., Müller, C., Seidl, H., Zalinescu, E.: Verifying security policies in multi-agent workflows with loops. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017. pp. 633–645. ACM (2017). https://doi.org/10.1145/3133956.3134080, `https://doi.org/10.1145/3133956.3134080`

34. Finkbeiner, B., Rabe, M.N., Sánchez, C.: Algorithms for model checking hyperltl and hyperctl ^*. In: Kroening, D., Pasareanu, C.S. (eds.) Computer Aided Verification - 27th International Conference, CAV 2015, San Francisco, CA, USA, July 18-24, 2015, Proceedings, Part I. Lecture Notes in Computer Science, vol. 9206, pp. 30–48. Springer (2015). https://doi.org/10.1007/978-3-319-21690-4_3, `https://doi.org/10.1007/978-3-319-21690-4_3`

35. Finkbeiner, B., Sipma, H.: Checking finite traces using alternating automata. Formal Methods Syst. Des. **24**(2), 101–127 (2004). https://doi.org/10.1023/B:FORM.0000017718.28096.48, `https://doi.org/10.1023/B:FORM.0000017718.28096.48`

36. Finkbeiner, B., Zimmermann, M.: The first-order logic of hyperproperties. In: 34th Symposium on Theoretical Aspects of Computer Science, STACS 2017, March 8-11, 2017, Hannover, Germany. LIPIcs, vol. 66, pp. 30:1–30:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2017). https://doi.org/10.4230/LIPIcs.STACS.2017.30, `https://doi.org/10.4230/LIPIcs.STACS.2017.30`

37. Fortin, M., Kuijer, L.B., Totzke, P., Zimmermann, M.: Hyperltl satisfiability is σ_1^1-complete, hyperctl* satisfiability is σ_1^2-complete. CoRR **abs/2105.04176** (2021), `https://arxiv.org/abs/2105.04176`

38. Gößler, G., Métayer, D.L.: A general trace-based framework of logical causality. In: Fiadeiro, J.L., Liu, Z., Xue, J. (eds.) Formal Aspects of Component Software - 10th International Symposium, FACS 2013, Nanchang, China, October 27-29, 2013, Revised Selected Papers. Lecture Notes in Computer Science, vol. 8348, pp. 157–173. Springer (2013). https://doi.org/10.1007/978-3-319-07602-7_11, `https://doi.org/10.1007/978-3-319-07602-7_11`

39. Gössler, G., Stefani, J.: Causality analysis and fault ascription in component-based systems. Theor. Comput. Sci. **837**, 158–180 (2020). https://doi.org/10.1016/j.tcs.2020.06.010, `https://doi.org/10.1016/j.tcs.2020.06.010`

40. Groce, A., Chaki, S., Kroening, D., Strichman, O.: Error explanation with distance metrics. Int. J. Softw. Tools Technol. Transf. **8**(3), 229–247 (2006). https://doi.org/10.1007/s10009-005-0202-0, `https://doi.org/10.1007/s10009-005-0202-0`

41. Groce, A., Kroening, D., Lerda, F.: Understanding counterexamples with explain. In: Alur, R., Peled, D.A. (eds.) Computer Aided Verification, 16th International Conference, CAV 2004, Boston, MA, USA, July 13-17, 2004,

Proceedings. Lecture Notes in Computer Science, vol. 3114, pp. 453–456. Springer (2004). https://doi.org/10.1007/978-3-540-27813-9_35, `https://doi.org/10.1007/978-3-540-27813-9_35`

42. Groce, A., Visser, W.: What went wrong: Explaining counterexamples. In: Ball, T., Rajamani, S.K. (eds.) Model Checking Software, 10th International SPIN Workshop. Portland, OR, USA, May 9-10, 2003, Proceedings. Lecture Notes in Computer Science, vol. 2648, pp. 121–135. Springer (2003). https://doi.org/10.1007/3-540-44829-2_8, `https://doi.org/10.1007/3-540-44829-2_8`

43. Halpern, J.Y.: A modification of the halpern-pearl definition of causality. In: Yang, Q., Wooldridge, M.J. (eds.) Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. pp. 3022–3033. AAAI Press (2015), `http://ijcai.org/Abstract/15/427`

44. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part i: Causes. The British Journal for the Philosophy of Science **56**(4), 843–887 (2005), `http://www.jstor.org/stable/3541870`

45. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part ii: Explanations. The British Journal for the Philosophy of Science **56**(4), 889–911 (2005), `http://www.jstor.org/stable/3541871`

46. Holzmann, G.J.: The model checker SPIN. IEEE Trans. Software Eng. **23**(5), 279–295 (1997). https://doi.org/10.1109/32.588521, `https://doi.org/10.1109/32.588521`

47. Horak, T., Coenen, N., Metzger, N., Hahn, C., Flemisch, T., Méndez, J., Dimov, D., Finkbeiner, B., Dachselt, R.: Visual analysis of hyperproperties for understanding model checking results. IEEE Trans. Vis. Comput. Graph. **28**(1), 357–367 (2022). https://doi.org/10.1109/TVCG.2021.3114866, `https://doi.org/10.1109/TVCG.2021.3114866`

48. Hsu, T., Sánchez, C., Bonakdarpour, B.: Bounded model checking for hyperproperties. In: Tools and Algorithms for the Construction and Analysis of Systems - 27th International Conference, TACAS 2021, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2021, Luxembourg City, Luxembourg, March 27 - April 1, 2021, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12651, pp. 94–112. Springer (2021). https://doi.org/10.1007/978-3-030-72016-2_6, `https://doi.org/10.1007/978-3-030-72016-2_6`

49. Ignatiev, A., Morgado, A., Marques-Silva, J.: PySAT: A Python toolkit for prototyping with SAT oracles. In: SAT. pp. 428–437 (2018), `https://doi.org/10.1007/978-3-319-94144-8_26`

50. Jee, E., Jeon, S., Cha, S.D., Koh, K.Y., Yoo, J., Park, G., Seong, P.: Fbdverifier: Interactive and visual analysis of counterexample in formal verification of function block diagram. J. Res. Pract. Inf. Technol. **42**(3), 171–188 (2010), `http://ws.acs.org.au/jrpit/JRPIT42.3.171.pdf`

51. Kocher, P., Horn, J., Fogh, A., , Genkin, D., Gruss, D., Haas, W., Hamburg, M., Lipp, M., Mangard, S., Prescher, T., Schwarz, M., Yarom, Y.: Spectre attacks: Exploiting speculative execution. In: 40th IEEE Symposium on Security and Privacy (S&P'19) (2019)

52. Krebs, A., Meier, A., Virtema, J., Zimmermann, M.: Team semantics for the specification and verification of hyperproperties. In: 43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August 27-31, 2018, Liverpool, UK. LIPIcs, vol. 117, pp. 10:1–10:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2018). https://doi.org/10.4230/LIPIcs.MFCS.2018.10, `https://doi.org/10.4230/LIPIcs.MFCS.2018.10`

53. Lahtinen, J., Launiainen, T., Heljanko, K., Ropponen, J.: Model checking methodology for large systems, faults and asynchronous behaviour: SARANA 2011 work report. No. 12 in VTT Technology, VTT Technical Research Centre of Finland, Finland (2012)
54. Larsen, K.G., Pettersson, P., Yi, W.: UPPAAL in a nutshell. Int. J. Softw. Tools Technol. Transf. **1**(1-2), 134–152 (1997), `https://doi.org/10.1007/s100090050010`
55. Leitner-Fischer, F., Leue, S.: Causality checking for complex system models. In: Giacobazzi, R., Berdine, J., Mastroeni, I. (eds.) Verification, Model Checking, and Abstract Interpretation, 14th International Conference, VMCAI 2013, Rome, Italy, January 20-22, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7737, pp. 248–267. Springer (2013). https://doi.org/10.1007/978-3-642-35873-9_16, `https://doi.org/10.1007/978-3-642-35873-9_16`
56. Leitner-Fischer, F., Leue, S.: Probabilistic fault tree synthesis using causality computation. Int. J. Crit. Comput. Based Syst. **4**(2), 119–143 (2013). https://doi.org/10.1504/IJCCBS.2013.056492, `https://doi.org/10.1504/IJCCBS.2013.056492`
57. Lewis, D.: Causation. Journal of Philosophy **70**(17), 556–567 (1973). https://doi.org/10.2307/2025310
58. Lipp, M., Schwarz, M., Gruss, D., Prescher, T., Haas, W., Fogh, A., Horn, J., Mangard, S., Kocher, P., Genkin, D., Yarom, Y., Hamburg, M.: Meltdown: Reading kernel memory from user space. In: 27th USENIX Security Symposium (USENIX Security 18) (2018)
59. M., V.C., Rabe, M.: py-aiger (version 6) [computer software]. `https://github.com/mvcisback/py-aiger`
60. Mascle, C., Zimmermann, M.: The keys to decidable hyperltl satisfiability: Small models or very simple formulas. In: 28th EACSL Annual Conference on Computer Science Logic, CSL 2020, January 13-16, 2020, Barcelona, Spain. LIPIcs, vol. 152, pp. 29:1–29:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2020). https://doi.org/10.4230/LIPIcs.CSL.2020.29, `https://doi.org/10.4230/LIPIcs.CSL.2020.29`
61. Moore, E.F., et al.: Gedanken-experiments on sequential machines. Automata studies **34**, 129–153 (1956)
62. Pakonen, A., Buzhinsky, I., Vyatkin, V.: Counterexample visualization and explanation for function block diagrams. In: 16th IEEE International Conference on Industrial Informatics, INDIN 2018, Porto, Portugal, July 18-20, 2018. pp. 747–753. IEEE (2018). https://doi.org/10.1109/INDIN.2018.8472025, `https://doi.org/10.1109/INDIN.2018.8472025`
63. Pnueli, A.: The temporal logic of programs. In: 18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977. pp. 46–57. IEEE Computer Society (1977). https://doi.org/10.1109/SFCS.1977.32, `https://doi.org/10.1109/SFCS.1977.32`
64. Schuppan, V., Biere, A.: Shortest counterexamples for symbolic model checking of LTL with past. In: Halbwachs, N., Zuck, L.D. (eds.) Tools and Algorithms for the Construction and Analysis of Systems, 11th International Conference, TACAS 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005, Edinburgh, UK, April 4-8, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3440, pp. 493–509. Springer (2005). https://doi.org/10.1007/978-3-540-31980-1_32, `https://doi.org/10.1007/978-3-540-31980-1_32`

65. Sörensson, N.: Minisat 2.2 and minisat++ 1.1. A short description in SAT Race **2010** (2010)
66. Stucki, S., Sánchez, C., Schneider, G., Bonakdarpour, B.: Gray-box monitoring of hyperproperties. In: Formal Methods - The Next 30 Years - Third World Congress, FM 2019, Porto, Portugal, October 7-11, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11800, pp. 406–424. Springer (2019). https://doi.org/10.1007/978-3-030-30942-8_25, `https://doi.org/10.1007/978-3-030-30942-8_25`
67. Vardi, M.Y.: Alternating automata: Unifying truth and validity checking for temporal logics. In: McCune, W. (ed.) Automated Deduction - CADE-14, 14th International Conference on Automated Deduction, Townsville, North Queensland, Australia, July 13-17, 1997, Proceedings. Lecture Notes in Computer Science, vol. 1249, pp. 191–206. Springer (1997). https://doi.org/10.1007/3-540-63104-6_19, `https://doi.org/10.1007/3-540-63104-6_19`
68. Wang, C., Yang, Z., Ivancic, F., Gupta, A.: Whodunit? causal analysis for counterexamples. In: Graf, S., Zhang, W. (eds.) Automated Technology for Verification and Analysis, 4th International Symposium, ATVA 2006, Beijing, China, October 23-26, 2006. Lecture Notes in Computer Science, vol. 4218, pp. 82–95. Springer (2006). https://doi.org/10.1007/11901914_9, `https://doi.org/10.1007/11901914_9`

A   Appendix

A.1   Alternating Automata

Alternating automata [67] are automata over infinite words that generalize nondeterministic and universal automata. An alternating automaton $\mathcal{A}$ is defined by the following grammar

$$\epsilon_{\mathcal{A}} \mid \langle \nu, \mathcal{A}, f \rangle \mid \mathcal{A} \wedge \mathcal{A} \mid \mathcal{A} \vee \mathcal{A}$$

where $\epsilon_{\mathcal{A}}$ is the empty automaton; $\mathcal{A} \wedge \mathcal{A}$ and $\mathcal{A} \vee \mathcal{A}$ are a disjunction and conjunction of two automata, respectively; and $n = \langle \nu, \mathcal{A}, f \rangle$ is a node, such that ν is a state formula that labels n, $\mathcal{A}$ is the next state (automaton), and f indicates whether n is accepting or rejecting (acc/rec). Since runs of an alternating automaton are defined using conjunctions, they form a *run tree* (see [67] for a formal definition), where disjunctions express non-determinism. The set of words accepted (using Büchi acceptance condition) by an alternating automaton are those who have a run such that all of its branches in the tree visit infinitely often in an accepting state. The language $\mathcal{L}(\mathcal{A})$ of an automaton $\mathcal{A}$ is the set of words accepted by $\mathcal{A}$.

The set of nodes of an alternating automaton $\mathcal{A}$ is denoted by $\mathcal{N}(\mathcal{A})$, where $\mathcal{N}(\epsilon_{\mathcal{A}}) = \emptyset$, $\mathcal{N}(\langle \nu, \mathcal{A}, f \rangle) = \mathcal{N}(\mathcal{A})$, and $\mathcal{N}(\mathcal{A} \wedge \mathcal{A}') = \mathcal{N}(\mathcal{A} \vee \mathcal{A}') = \mathcal{N}(\mathcal{A}) \cup \mathcal{N}(\mathcal{A}')$. For an LTL formula φ there is a linear translation to an alternating automaton $\mathcal{A}_{\varphi}$, s.t. $\mathcal{L}(\mathcal{A}_{\varphi})$ is the set of traces that satisfy φ [67]. In this construction, $\mathcal{N}(\mathcal{A})$ is the set of subformulas of φ and their negations, $\mathcal{A}$ is φ, and all formulas of the form $\neg(\varphi_1 \, \mathcal{U} \, \varphi_2)$ are *accepting*.

A.2   LTL to Alternating Automata

For LTL formulas φ and φ' and $a \in AP$

$$\mathcal{A}(a) = \langle a, \epsilon_{\mathcal{A}}, acc \rangle$$
$$\mathcal{A}(\neg a) = \langle \neg a, \epsilon_{\mathcal{A}}, acc \rangle$$
$$\mathcal{A}(\varphi \wedge \varphi') = \mathcal{A}(\varphi) \wedge \mathcal{A}(\varphi')$$
$$\mathcal{A}(\varphi \vee \varphi') = \mathcal{A}(\varphi) \vee \mathcal{A}(\varphi')$$
$$\mathcal{A}(\bigcirc \varphi) = \langle true, \mathcal{A}(\varphi), rej \rangle$$
$$\mathcal{A}(\square \varphi) = \langle true, \mathcal{A}(\square \varphi), acc \rangle \wedge \mathcal{A}(\varphi)$$
$$\mathcal{A}(\diamondsuit \varphi) = \langle true, \mathcal{A}(\diamondsuit \varphi), rej \rangle \wedge \mathcal{A}(\varphi)$$
$$\mathcal{A}(\varphi \mathcal{U} \varphi') = \mathcal{A}(\varphi') \vee (\langle true, \mathcal{A}(\varphi \mathcal{U} \varphi'), rej \rangle \wedge \mathcal{A}(\varphi))$$
$$\mathcal{A}(\varphi \mathcal{R} \varphi') = (\mathcal{A}(\varphi) \wedge \mathcal{A}(\varphi')) \vee (\langle true, \mathcal{A}(\varphi \mathcal{R} \varphi'), acc \rangle \wedge \mathcal{A}(\varphi'))$$

A.3   Counterfactual Automaton for Running Example

The counterfactual automaton $T_{t_2}^C$ for system T (see Fig. 1) and the trace $t_2 = \{hi\}\{hi, ho\}\{ho, lo\}^{\omega}$ from the running example (see Sec. 3) is illustrated in Fig. 2. For readability we only pictured the reachable fragment of the automaton, the states $(s_1, 0), (s_2, 0)$ and $(s_3, 0)$ are missing. Note that following t_2 can be done in $T_{t_2}^C$ without the contingency variables, e.g., lo^C. If we diverge from the input sequence of t_2, however, such as when taking an edge labeled with $\neg hi$ in the first step, then we can use the contingency variables to set the corresponding value as in t_2. For instance, the edge labeled with $\neg hi \wedge lo^C \wedge \neg ho^C$ goes to $(s_0, 1)$ because we remove lo from the outputs of the successor state. We then proceed as in the state with the corresponding label, i.e., s_0. Automata copies in the chain that correspond to the prefix of the trace cannot be visited twice, for instance in $T_{t_2}^C$ we are forced to loop in the third copy and can visit the first two only once.

A.4   Cause Candidates for Asymmetric Arbiter

In the following, we provide the cause candidates $\tilde{\mathcal{C}}$, which are given as an intermediate output of our implementation for both asymmetric arbiter instances. Especially in the `Asymmetric arbiter` instance, that tb_secret does not hold on timestep 3 on trace t_1, but does hold on timestep 3 on trace t_2 immediately catches the eye. The events are given as a tuple of the valuation of an input and its timestep, sorted by their trace belonging.
$\tilde{\mathcal{C}}$ for `Asymmetric arbiter`:

```
T1: [(!req_1, 6), (!req_0, 1), (req_1, 0), (req_0, 0),
     (req_0, 5), (req_1, 4), (!req_1, 2), (!req_0, 2),
     (!tb_secret, 3), (req_0, 3), (req_1', 3)]
T2: [(!req_0, 4), (!req_0, 1), (req_1, 0), (req_0, 0),
     (!req_0, 6), (!req_1, 6), (req_0, 5), (req_1, 5),
     (req_0, 3), (tb_secret, 3), (req_1, 3), (!req_1, 2),
     (!req_0, 2)]
```
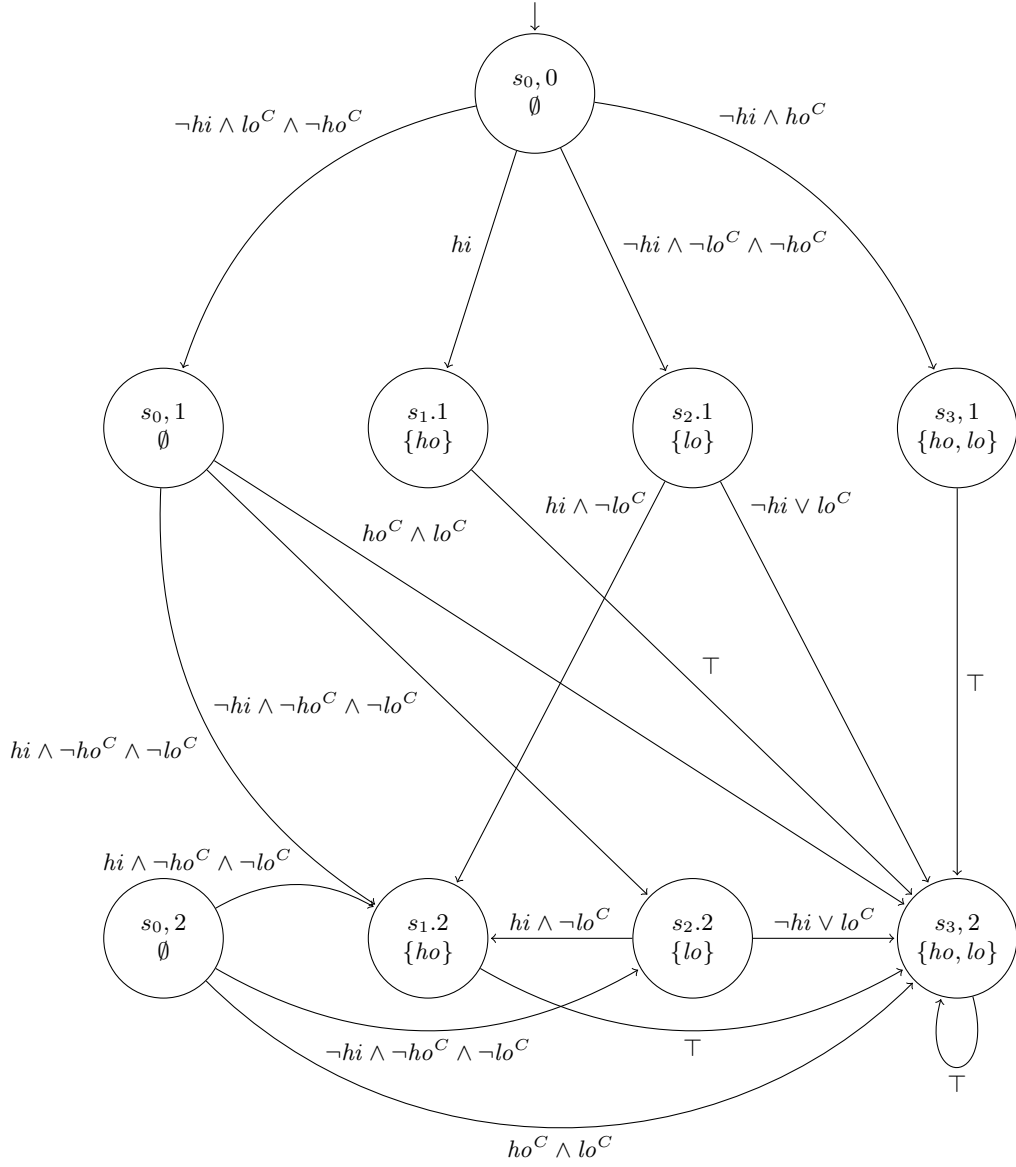
Fig. 2: The counterfactual automaton $T^C_{t_2}$ described in App. A.3.

$\tilde{\mathcal{C}}$ for `Asymmetric arbiter '19`:

T1: $[(\,!\,\mathrm{req0}\,,\ 1)\,,\ (\,\mathrm{req0}\,,\ 0)\,,\ (\,\mathrm{req1}\,,\ 0)\,,\ (\,!\,\mathrm{req0}\,,\ 2)\,,$
    $(\,!\,\mathrm{req1}\,,\ 2)\,]$
T2: $[(\,!\,\mathrm{req0\,'}\,,\ 1)\,,\ (\,\mathrm{req0}\,,\ 0)\,,\ (\,\mathrm{req1}\,,\ 0)\,,\ (\,!\,\mathrm{req0}\,,\ 2)\,,$
    $(\,!\,\mathrm{req1}\,,\ 2)\,]$

A.5   Specifications for Experimental Results

We used the following HyperLTL specifications given in MCHyper syntax on the respective benchmarks.
`Running Example:`

```
Forall (Forall (G (Eq (AP \"lo\" 0) (AP \"lo\" 1))))
```

`Security in & out:`

```
Forall (Forall (Implies (G (Eq (AP \"li\" 0) (AP \"li\"
    1))) (G (Eq (AP \"lo\" 0) (AP \"lo\" 1)))))
```

`Drone example 1:`

```
Forall (Forall (X (G (Implies (Eq (AP \"bound\" 0) (AP
    \"bound\" 1)) (X (Eq (AP \"emergency\" 0) (AP
    \"emergency\" 1)))))))
```

`Drone example 2:`

```
Forall (Forall (Or (G (Eq (AP \"increase\" 0) (AP
    \"increase\" 1))) (Until (Eq (AP \"increase\" 0) (AP
    \"increase\" 1)) (And (Eq (AP \"increase\" 0) (AP
    \"increase\" 1)) (Neq (AP \"up\" 0) (AP \"up\" 1))))))
```

`Asymmetric arbiter '19:`

```
Forall (Forall (Implies (G (And (Eq (AP \"req0\" 0) (AP
    \"req1\" 1)) (Eq (AP \"req1\" 0) (AP \"req0\" 1))))
    (G (And (Eq (AP \"grant0\" 0) (AP \"grant1\" 1)) (Eq
    (AP \"grant1\" 0) (AP \"grant0\" 1))))))
```

`Asymmetric arbiter:`

```
Forall (Forall (Implies (G (And (Eq (AP \"req_0\" 0) (AP
    \"req_0\" 1)) (Eq (AP \"req_1\" 0) (AP \"req_1\"
    1)))) (G (And (Eq (AP \"grant_0\" 0) (AP \"grant_0\"
    1)) (Eq (AP \"grant_1\" 0) (AP \"grant_1\" 1))))))
```