Counterfactual Causality for Reachability and Safety based on Distance Functions

Julie Parreaux
Aix Marseille Univ, CNRS, LIS, Marseille, France
julie.parreaux@univ-amu.fr

Jakob Piribauer
Technische Universität Dresden, Germany
Technische Universität München, Germany
jakob.piribauer@tu-dresden.de

Christel Baier
Technische Universität Dresden, Germany
christel.baier@tu-dresden.de

Investigations of causality in operational systems aim at providing human-understandable explanations of *why* a system behaves as it does. There is, in particular, a demand to explain what went wrong on a given counterexample execution that shows that a system does not satisfy a given specification. To this end, this paper investigates a notion of counterfactual causality in transition systems based on Stalnaker's and Lewis' semantics of counterfactuals in terms of most similar possible worlds and introduces a novel corresponding notion of counterfactual causality in two-player games. Using distance functions between paths in transition systems to capture the similarity of executions, this notion defines whether reaching a certain set of states is a cause for the fact that a given execution of a system satisfies an undesirable reachability or safety property. Similarly, using distance functions between memoryless strategies in reachability and safety games, it is defined whether reaching a set of states is a cause for the fact that a given strategy for the player under investigation is losing.

The contribution of the paper is two-fold: In transition systems, it is shown that counterfactual causality can be checked in polynomial time for three prominent distance functions between paths. In two-player games, the introduced notion of counterfactual causality is shown to be checkable in polynomial time for two natural distance functions between memoryless strategies. Further, a notion of explanation that can be extracted from a counterfactual cause and that pinpoints changes to be made to the given strategy in order to transform it into a winning strategy is defined. For the two distance functions under consideration, the problem to decide whether such an explanation imposes only minimal necessary changes to the given strategy with respect to the used distance function turns out to be coNP-complete and not to be solvable in polynomial time if P is not equal to NP, respectively.

1 Introduction

Modern software and hardware systems have reached a level of complexity that makes it impossible for humans to assess whether a system behaves as intended without tools tailored for this task. To tackle this problem, automated verification techniques have been developed. *Model checking* is one prominent such technique: A model-checking algorithm takes a mathematical model of the system under investigation and a formal specification of the intended behavior and determines whether all possible executions of the model satisfy the specification. While the results of a model-checking algorithm provide guarantees on

Funding: This work was partly funded by DFG Grant 389792660 as part of TRR 248 (Foundations of Perspicuous Software Systems), the Cluster of Excellence EXC 2050/1 (CeTI, project ID 390696704, as part of Germany's Excellence Strategy), and the DFG projects BA-1679/11-1 and BA-1679/12-1, and the ANR project Ticktac (ANR-18-CE40-0015).

A. Achilleos and D. Della Monica (Eds.): Fourteenth International Symposium on Games, Automata, Logics, and Formal Verification (GandALF 2023).
EPTCS 390, 2023, pp. 132–149, doi:10.4204/EPTCS.390.9

the correctness of a system or affirm the presence of an error, their usefulness is, nevertheless, limited as they do not provide a human-understandable explanation of the behavior of the system.

To provide additional information on *why* the system behaves as it does, certificates witnessing the result of the model-checking procedure, in particular counterexample traces in case of a negative result, have been studied extensively (see, e.g., [10, 29, 9, 30]). Due to the potentially still enormous size of counterexample traces and other certificates, a line of research has emerged that tries to distill comprehensible explications of what causes the system to behave as it does using formalizations of *causality* (see, e.g., [32, 33, 2]).

Forward- and backward-looking causality There are two fundamentally different types of notions of causality: *forward-looking* and *backward-looking* notions [34]. In the context of operational system models, forward-looking causality describes general causal relations between events that might happen along some possible executions. Backward-looking causality, on the other hand, addresses the causal relation between events along a given execution of the system model. This distinction is captured in more general contexts by the distinction between *type-level* causality addressing general causal dependencies between events that might happen when looking forward in a world model, and *token-level* or *actual* causality, corresponding to the backward view, that addresses causes for a particular event that actually happened (see, e.g., [16]).

Notions of *necessary* causality are typically forward-looking: A necessary cause C for an effect E is an event that occurs on every execution that exhibits the effect E (see, e.g., [3], and for a philosophical analysis of necessity in causes [28]). The backward view naturally arises when the task is to explain what went wrong after an undesired effect has been observed. In the verification context, the backward view is natural for explaining counterexamples, see e.g. [42, 6, 15, 35, 38, 39]. Most of these techniques rely on the *counterfactuality* principle, which has been originally studied in philosophy [20, 21, 37, 26, 27] and formalized mathematically by Halpern and Pearl [17, 18, 19, 16]. Intuitively, counterfactual causality requires that the effect would not have happened, if the cause had not occurred, in combination with some minimality constraints for causes. The most prominent account for the semantics of the involved counterfactual implication is provided by Stalnaker and Lewis [37, 26, 27] in terms of closest, i.e., most similar, possible worlds. The statement "if the cause C had not occurred, then the effect E would not have occurred" holds true if in the worlds that are most similar to the actual world and in which C did not occur, E also did not occur. Interpreting executions of a system as possible worlds, the actual world is an execution π where both the effect E and its counterfactual cause C occur, while the effect E does not occur in alternative executions that are as similar as possible to π and that do not exhibit C.

For a more detailed discussion on the distinction between forward- and backward looking causality and related concepts for responsibility, we refer the reader, e.g., to [34, 40, 41, 4, 2].

Defining counterfactual causality in transition systems and reachability games To define our backward-looking notion of counterfactual causality in transition systems, we follow an approach similar to the one by Groce et al [14] who presented a Stalnaker-Lewis-style formalization of counterfactual dependence of events using distance functions. We consider the case where effects are reachability or safety properties and causes are sets of states. To illustrate the idea, let $\mathscr{T}$ be a transition system and let E and C be disjoint sets of states of $\mathscr{T}$ indicating a reachability effect and a potential cause, respectively. Consider an execution π that reaches the effect set and the potential cause set. We employ the counterfactual reading of causality by Stalnaker and Lewis by viewing executions as possible worlds using a similarity metric d on paths: Reaching C was a cause for π to reach E if all paths ζ, that do not

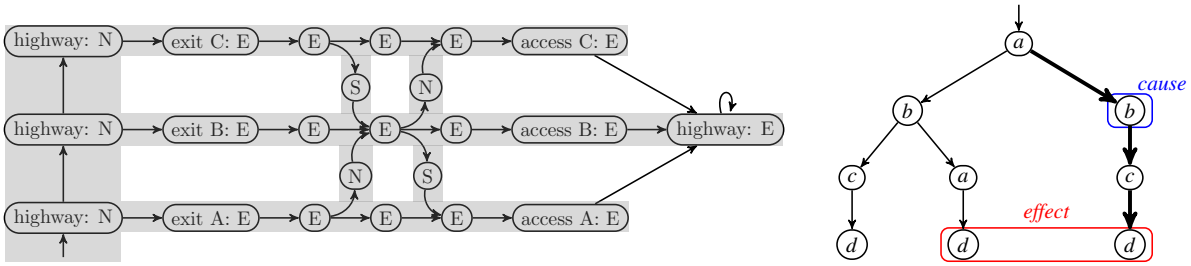

Figure 1: On the left: example transition system modelling a traffic grid (Ex. 1). On the right: Example of a d_{Hamm}-counterfactual cause that is not a d_{pref}-counterfactual cause (Ex. 3).

reach C and that are most similar to π according to d among all paths with this property, satisfy $\zeta \vDash \Box\neg E$, i.e., they do not reach E. So, we first determine the minimal similarity-distance $d_{\min} = \min\{d(\pi,\zeta) \mid \zeta \vDash \Box\neg C\}$ from π to a path ζ that does not reach C. Then, we check whether all paths that do not reach C and have similarity-distance $d_{\min}$ to π do not reach E: Do all $\zeta \in \{\zeta' \mid d(\pi,\zeta') = d_{\min} \text{ and } \zeta' \vDash \Box\neg C\}$ satisfy $\Box\neg E$? If the answer is yes, it is the case that "if C had not occurred, then E would not have occurred" and so C is a counterfactual cause for E on π.

Example 1. Consider the following distance function on paths in a labeled transition system $\mathscr{T}$ with states S and a labeling function $L\colon S \to \mathscr{A}$ for a set of labels $\mathscr{A}$: For paths $\pi = s_0, s_1, \ldots$ and $\pi' = t_0, t_1, \ldots$, we define $dist(\pi,\pi') = |\{n \in \mathbb{N} \mid L(s_n) \neq L(t_n)\}|$. So, paths are more similar if their traces differ at fewer positions. To determine whether C is a cause for E on π, we first determine what the least number $n_{\min}$ of changes to the state labels of π is to obtain a path ζ that does not reach C. Then, we have to check whether all paths differing from π in $n_{\min}$ labels and not reaching C do not reach E. If this is the case, C is a counterfactual cause for E on π with respect to $dist$.

Now, consider the example transition system $\mathscr{T}$ modelling a road system with a highway going north that has three exits into a small town which can be left again on a highway heading east depicted in Fig. 1. Each state is labeled with N, E, or S for north, east, and south as indicated in Fig. 1 depending on the direction the cars move on the respective road. Say, an agent traverses the system via the path π with trace NNE^ω, i.e., by taking exit B from the first highway and then going eastwards straight through the town. Assume that there is a traffic jam on access B while the other access roads are free. The question is now whether taking exit B was a cause for being stuck in slow traffic later on, i.e., for the effect $\{\text{access B}\}$. First, note here that the set $\{\text{exit B}\}$ is not a forward-looking necessary cause for reaching $\{\text{access B}\}$. There are paths through the system that avoid $\{\text{exit B}\}$, but reach $\{\text{access B}\}$.

However, given the fact that the agent traversed the town by going straight eastwards, it is reasonable to say that the agent would have reached a different access road if she had taken a different exit from the first highway. This is reflected in the counterfactual definition using $dist$: There are two paths that do not reach exit B and whose traces differ from π at only one position, namely the paths with trace NE^ω and $NNNE^\omega$. These paths do also not reach access B. So, $\{\text{exit B}\}$ is a counterfactual cause for $\{\text{access B}\}$ on π; if the agent had taken exit A or C, she would not have hit the low traffic flow at access B. ⌟

In the context of two-player reachability games, causality has been used as a tool to solve games [1]. In our work, we focus on explaining why a certain strategy does not allow the player to win. More precisely, in reachability games between players with a safety and the complementing reachability objective, respectively, we consider the situation where one of the players Π has a winning strategy, but loses the game using a strategy σ. We introduce a notion of counterfactual causality that aims to provide insights into what is wrong with strategy σ by transferring the counterfactual definition using distance functions

distance d	causality
prefix	in P (Thm. 4)
Hamming	in P (Thm. 5)
Levenshtein	in P (Thm. 8)

distance d	causality	explanations
Hausdorff lifting d^H_{pref} of the prefix distance	in P (Thm. 12)	
Hamming strategy distance d^s_{Hamm}	in P in acyclic games (Thm. 13)	coNP-complete (Cor. 21)
Hausdorff-inspired distance d^*		not in P if P≠NP (Cor. 21)

Table 1: Overview of the complexity results. On the left, the complexities of checking d-counterfactual causality in transition systems, and on the right, the complexities of checking d-counterfactual causality and d-minimality of explanations in reachability games.

d on memoryless strategies. A set of states C is said to be a d-counterfactual cause for the fact that σ is losing if all memoryless strategies τ, that make sure that C is not reached and have minimal d-distance to σ among all such strategies, are winning. Furthermore, we introduce *counterfactual explanations* that specify minimally invasive changes of σ's decisions required to turn σ into a winning strategy.

Contributions

- We show that d-counterfactual causal relationships in transition systems (defined as in [14]) can be checked in polynomial time for the following three distance metrics d (Sec. 3.2):
  1. the prefix distance: paths are more similar if their traces share a longer prefix.
  2. the Hamming distance that counts the positions at which traces of paths differ.
  3. the Levenshtein distance that counts how many insertions, deletions, and substitutions are necessary to transform the trace of one path to the trace of another path.

  Furthermore, we show that the notion of d-counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes [18, 19].
- In reachability games, we provide a generalization of this notion using similarity metrics on memoryless deterministic strategies. We show that for the Hausdorff lifting of the prefix distance on paths to a distance function on memoryless deterministic strategies, the resulting notion can be checked in polynomial time (Sec. 4.1).
- We introduce a notion of *counterfactual explanation* that can be computed from a counterfactual cause (Sec. 4.2). An explanation specifies where a non-winning strategy needs to be changed. Of particular interest are D-minimal explanations that enforce only minimal necessary changes with respect to a distance function D on strategies. For two distance functions related to the Hamming distance, we show that checking whether an explanation is minimal is coNP-complete and not in P if P≠NP, respectively.

An overview of the complexity results can be found in Table 1.

Related work Ways to pinpoint the problematic steps in a counterexample trace by localizing errors have widely been studied [42, 6, 15, 35, 38, 39]. For counterfactuality in transition systems, we follow the approach of [14] with distance metrics. In contrast to the causes in this paper, causes in [14] are formulas in an expressive logic that can precisely talk about the valuation of variables after a certain number of steps. Further [14] is not concerned with checking causality, but with finding causes, which,

due to the expressive type of causes, algorithmically boils down to finding executions avoiding the effect with a minimal distance to the given one.

Based on counterfactuality, Halpern and Pearl [18, 19, 16] provided an influential formalization of causality using structural equation models, which has served as the basis for various notions of causality in the verification context (see,e.g., [7, 24]). A key ingredient is the notion of *intervention* to provide a semantics for the counterfactual implication in Hume's definition of causality. An intervention in a structural equation model sets a variable to a certain value by force, ignoring its dependencies on other variables, and evaluates the effects of this enforced change. In a sense, a minimal set of interventions to avoid a cause then leads to a most similar execution avoiding the cause. We will discuss the relations between our definition and the Halpern-Pearl definition in more detail in Section 3.3. In [7], interventions are employed to counterexample traces in transition systems by allowing to flip atomic propositions along a trace. In contrast to our notion of counterfactual causes, this is tailored for complex linear time properties, but does not provide insights for reachability and safety. Furthermore, the flipping of atomic propositions can be seen as a change in the transition system while our definition considers alternative executions without manipulating the system. In [11], the Halpern-Pearl approach is applied to provide a counterfactual definition of causality in reactive systems. A distance partial order, namely the subset relation on sets of positions at which traces differ, is used to describe which interventions are acceptable as they constitute minimal changes necessary to avoid the cause. Checking causality is shown to be decidable by a formulation as a hyperlogic model-checking problem. Furthermore, notions of necessary and sufficient causes as sets of states in transition systems have been considered [3]. These do not rely on the counterfactuality principle and are of forward-looking nature.

We are not aware of formalisations of causality in game structures. The related concept of responsibility, has been investigated in multi-agent models [40, 41]. Notions of forward and backward responsibility of players in multi-player game structures with acyclic tree-like arena have been studied [4].

For a detailed overview of work on causality and related concepts in operational models, we refer the reader to the survey articles [8, 2].

2 Preliminaries

We briefly present notions we use and our notation. For details, see [5, 13].

Transition systems. A transition system is a tuple $\mathscr{T} = (S, s_{init}, \to, L)$ where S is a finite set of states, $s_{init} \in S$ is an initial state, $\to \subseteq S \times S$ is a transition relation and $L\colon S \to 2^{\mathsf{AP}}$ is a labeling function where AP is a set of atomic propositions. A path in a transition system is a finite or infinite sequence of states $s_0 s_1 \dots$ such that $s_0 = s_{init}$ and, for all suitable indices i, there is a transition from s_i to s_{i+1}, i.e., $(s_i, s_{i+1}) \in \to$. Given a path $\pi = s_0 s_1 \dots$, we denote its trace $L(s_0)L(s_1)\dots$ by $L(\pi)$. If there are no outgoing transitions from a state, we call the state *terminal*.

Computation tree logic (CTL). The branching-time logic CTL consists of state formulas that are evaluated at states in a transition system formed by $\Phi ::= \top \mid a \mid \Phi \wedge \Phi \mid \neg\Phi \mid \exists\varphi \mid \forall\varphi$ where $a \in \mathsf{AP}$ is an atomic proposition and path formulas evaluated on paths formed by $\varphi ::= \bigcirc\Phi \mid \Phi \mathrm{U} \Phi$. The semantics of the temporal operators in path formulas is as usual. We use the abbreviations $\Diamond\Phi$ for $\top \mathrm{U} \Phi$ and $\Box\Phi = \neg\Diamond\neg\Phi$ and also allow sets of states T in the place of state formulas. The semantics of $\exists\varphi$ are that there exists a path starting in the state at which the formula is evaluated that satisfies φ; $\forall\varphi$ is defined dually to that as usual. Model checking of CTL-formulas can be done in polynomial time. For details, see [5].

Reachability games. A *reachability game* is a tuple $\mathscr{G} = (V, v_{\mathsf{i}}, \Delta)$ where $V = V_{\mathsf{Reach}} \uplus V_{\mathsf{Safe}} \uplus V_{Eff}$ is the set of vertices shared between players Reach and Safe, and some target vertices V_{Eff} (*Eff* for effect).

$v_\mathsf{i} \in V \setminus V_{Eff}$ is the initial vertex and $\Delta \subseteq V \times V$ is the set of edges. We denote by $\Delta(v)$ the set of edges from v. W.l.o.g., we assume that target vertices are terminal states, i.e. for all vertices $v \in V_{Eff}$, $\Delta(v) = \emptyset$. A *finite play* is a finite sequence of vertices $\pi = v_0 v_1 \cdots v_k \in V^*$ such that for all $0 \leq i < k$, $(v_i, v_{i+1}) \in \Delta$. A *play* is either a finite play ending in a target vertex, or an infinite sequence of vertices such that every finite prefix is a finite play. Transition systems can be viewed as one-player games.

A *strategy* for Reach in a reachability game $\mathscr{G}$ is a mapping $\sigma \colon V^* V_{\mathsf{Reach}} \to V$. A play or finite play $\pi = v_0 v_1 \cdots$ is a σ-*play* if for all k with $v_k \in V_{\mathsf{Reach}}$, we have $\sigma(v_0 \cdots v_k) = v_{k+1}$. A strategy σ is an *MD-strategy* (for memoryless deterministic) if for all finite plays ξ and ξ' with the same last vertex, we have that $\sigma(\xi) = \sigma(\xi')$. In this paper, we mainly use MD-strategies and write $\sigma(v_k)$ instead of $\sigma(v_0 \cdots v_k)$ for MD-strategies σ. Moreover, under a (partial) MD-strategy σ, we define the *reachability game under* σ, denoted by $\mathscr{G}^\sigma = (V, v_\mathsf{i}, \Delta^\sigma)$, by removing edges not chosen by σ, i.e., $\Delta^\sigma = \Delta \setminus \{(v, v') \in \Delta \mid v \in V_{\mathsf{Reach}} \text{ and } \sigma(v) \text{ is defined and } \sigma(v) \neq (v, v')\}$. When σ is completely defined, $\mathscr{G}^\sigma$ is a transition system. Finally, a strategy is *winning* if all σ-plays starting in v_i end in a target vertex. Analogous definitions apply to Safe. In reachability games, either Reach or Safe wins with an MD-strategy. This winning strategy can be computed in polynomial time (see, e.g., [13]).

Distance function. A *distance function* on a set A is a function $d \colon A \times A \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that $d(x,x) = 0$ for all $x \in A$ and $d(x,y) = d(y,x)$ for all $x, y \in A$. It is called a *pseudo-metric* if additionally $d(x,y) + d(y,z) \geq d(x,z)$ for all $x, y, z \in A$, and a *metric* if further $d(x,y) = 0$ holds iff $x = y$ for all $x, y \in A$.

3 Counterfactual causes in transition systems

In this section, we introduce the backward-looking notion of counterfactual causes in transition systems using distance functions (Section 3.1). Afterwards, we prove that the definition can be checked in polynomial time for three well-known distance functions (Section 3.2). Finally, we illustrate similarities between our notion of counterfactual causality to the definition of causality by Halpern and Pearl (Sec 3.3). Proofs omitted here can be found in the extended version [31].

3.1 Definition

The effects we consider are reachability or safety properties $\Phi = \Diamond E$ or $\Phi = \Box \neg E$ for a set of states E. As the behavior of the system after E has been seen is not relevant for these properties, we assume that E consists of terminal states.

Definition 2 (d-counterfactual cause in transition systems)**.** Let $\mathscr{T}$ be a transition system and let d be a distance function on the set of maximal paths of $\mathscr{T}$. Let E be a set of terminal states and let C be a set of states disjoint from E. Let $\Phi = \Diamond E$ or $\Phi = \Box \neg E$. Given a maximal path π that visits C and satisfies Φ in $\mathscr{T}$, we say that C is a *d-counterfactual cause for Φ on π* if

1. there is a maximal path ρ in $\mathscr{T}$ that does not visit C, and
2. all maximal paths ρ with $\rho \vDash \Box \neg C$ with minimal distance to π do not satisfy Φ. In other words, all maximal paths ρ with $\rho \vDash \Box \neg C$ such that $d(\pi, \rho) \leq d(\pi, \rho')$ for all ρ' with $\rho' \vDash \Box \neg C$ satisfy $\rho \vDash \neg \Phi$.

The choice of the similarity distance d of course heavily influences the notion of d-counterfactual cause. In this paper, we will instantiate the definition with three distance functions that are among the most prominent distance functions between traces (or words). An experimental investigation to clarify in which situations what kind of distance functions leads to a desirable notion of causality, however, remains as future work.

Prefix metrics d^{AP}_{pref} and d_{pref}: given two paths π and ρ, let $n(\pi,\rho)$ be the length of the longest common prefix of their traces $L(\pi)$ and $L(\rho)$. Then, $d^{\mathsf{AP}}_{pref}(\pi,\rho) \stackrel{\text{def}}{=} 2^{-n(\pi,\rho)}$. We can also define the distance on paths instead of traces, which will be used later on: $d_{pref}(\pi,\rho) \stackrel{\text{def}}{=} 2^{-m(\pi,\rho)}$ where $m(\pi,\rho)$ is the length of the longest common prefix of π and ρ as paths. This can be seen as a special case of d^{AP}_{pref} if we assume that all states have a unique label.

The prefix metric measures similarity in a temporal way saying that executions are more similar if they initially agree for a longer period of time. If no further structure of the transition system or meaning of the labels is known, this distance function might be a reasonable choice for counterfactual causality.

Hamming distance d_{Hamm}: Given two words $w = w_0 \dots w_n$ and $v = v_0 \dots v_n$ of the same length, we define $d_{Hamm}(w,v) \stackrel{\text{def}}{=} |\{0 \leq i \leq n \mid w_i \neq v_i\}|$. For two maximal paths π and ρ of the same length in a transition system $\mathscr{T}$ with labeling function L, we define $d_{Hamm}(\pi,\rho) \stackrel{\text{def}}{=} d_{Hamm}(L(\pi),L(\rho))$. So, the distance between two paths is the Hamming distance of their traces.

The Hamming distance seems to be a reasonable measure if a system naturally proceeds through different layers, e.g., if a counter is increased in each step. Then, traces are viewed to be more similar if they agree on more layers. The temporal order of these layers, however, does not play a role.

Levenshtein distance d_{Lev} [25]: Given two words $w = w_0 \dots w_n$ and $v = v_0 \dots v_m$, the Levenshtein distance is defined as the minimal number of editing operations needed to produce v from w where the allowed operations are insertion of a letter, deletion of a letter, and substitution of a letter by a different letter. Formally, we define d_{Lev} in terms of *edit sequences*. Let Σ be an alphabet and $v,w \in \Sigma^* \cup \Sigma^\omega$ be two words over Σ. The *edit alphabet* for Σ is defined as $\Gamma \stackrel{\text{def}}{=} (\Sigma \cup \{\varepsilon\})^2 \setminus \{(\varepsilon,\varepsilon)\}$ where ε is a fresh symbol. An edit sequence for v and w is now a word $\gamma \in \Gamma^* \cup \Gamma^\omega$ such that the projection of γ onto the first component results in v when all εs are removed and the projection of γ onto the second component results in w when all εs are removed. E.g., let $\Sigma = \{a,b,c\}$, $v = abbc$ and $w = accbc$. One edit sequence is $\gamma = (a,a)(b,c)(\varepsilon,c)(b,b)(c,c)$. The weight of an edit sequence $\gamma = \gamma_1\gamma_2\dots$ is defined as $wgt(\gamma) = |\{i \mid \gamma_i \neq (\sigma,\sigma) \text{ for all } \sigma \in \Sigma\}|$. Then, for all words $v \in \Sigma^* \cup \Sigma^\omega$ and $w \in \Sigma^* \cup \Sigma^\omega$, we define $d_{Lev}(v,w) = \min\{wgt(\gamma) \mid \gamma \text{ is an edit sequence for } v \text{ and } w\}$. Again, we obtain a pseudo-metric on paths via the Levenshtein metric on traces.

The Levenshtein distance might be particularly useful if labels model actions that are taken. Two executions that are obtained by sequences of actions that only differ by inserting or leaving out some actions, but otherwise using the same actions, are considered to be similar in this case.

Example 3. Let us illustrate counterfactual causality for the prefix metric d_{pref} and the Hamming distance d_{Hamm}. Consider the transition system depicted in Figure 1. A path π as indicated by the bold arrows on the right via the potential cause to the effect has been taken: This is not a d_{pref}-counterfactual cause on π: The most similar paths to π that do not reach *cause* are both paths that move to the left initially. As one of these paths reaches *effect*, the set *cause* is not a d_{pref}-counterfactual cause for reaching *effect*.

Considering the distance function d_{Hamm} with the labels of the states as in Figure 1, we get a different result: The trace of π is *abcd*. The paths that avoid the potential cause have traces *abcd* and *abad*, respectively. So, the most similar path avoiding *cause* is the path on the left with trace *abcd* that also avoids *effect*. So, *cause* is a d_{Hamm}-counterfactual cause on π for $\Diamond$*effect*. Intuitively, this can be understood as saying if the system had avoided *cause* but otherwise behaved (as similar as possible to) as it did in terms of the produced trace, the effect would not have occurred. In particular, if labels represent actions that have been chosen, this is a reasonable reading of causality. ⌟

3.2 Checking counterfactual causality in transition systems

In this section, we provide algorithms to check d-counterfactual causality for the three distance functions d_{pref}^{AP}, d_{Hamm}, and d_{Lev}. For these algorithms, a maximal execution π of the system has to be given. We assume that π is a finite path ending in a terminal state. The problem to find causes that are small or satisfy other desirable properties is not addressed in this paper and remains as future work. We will briefly come back to this in the conclusions.

Prefix distance. First, we consider d_{pref}^{AP}-counterfactual causality and hence d_{pref}-counterfactual causality as a special case.

Theorem 4. *Let $\mathscr{T} = (S, s_{init}, \to, L)$ be a transition system, E a set of terminal states, C a set of states disjoint from E, and $\Phi = \Diamond E$ or $\Phi = \Box\neg E$. Let $\pi = s_0 \dots s_n$ be an execution reaching C and satisfying Φ. It is decidable in polynomial time whether C is a d_{pref}^{AP}-counterfactual cause for Φ on π.*

Proof sketch. The following algorithm solves the problem in polynomial time: First, we determine the last index i s.t. C is not reached on any path with trace $L(s_0), \dots, L(s_i)$ and s.t. C is avoidable from some state that is reachable via a path with trace $L(s_0), \dots, L(s_i)$. In order to that, we recursively construct sets T_{j+1} of states that are reachable via paths with trace $L(s_0), \dots, L(s_{j+1})$ and check for all states $t \in T_{j+1}$ whether $t \vDash \exists\Box\neg C$. If no such state exists, we have found the first index $j+1$ such that C is not avoidable anymore after trace $L(s_0), \dots, L(s_{j+1})$; so we have found $i = j$. Now, we check whether $t \vDash \forall(\Phi \to \Diamond C)$ for all $t \in T_i$. If this is the case, C is a d_{pref}^{AP}-counterfactual cause for E on π; otherwise, it is not. □

Hamming distance. The Hamming distance is only defined for words of the same length. We will hence first consider only transition systems in which all maximal paths have the same length. We can think of such transition systems as being structured in layers with indices 1 to k for some k. Transitions can then only move from a state on layer $i < k$ to a state on layer $i+1$. Afterwards, we consider a simple generalization of the Hamming distance to words of different lengths.
Original Hamming distance. Let $\mathscr{T} = (S, s_{init}, \to, L)$ be a transition system in which all maximal paths have the same length k. We annotate all states with the layer they are on: For each state $s \in S$, there is a unique length $n \leq k$ of all paths from s_{init} to s. We will say that state s lies on layer n in this case. By our assumption that effect states are terminal, the states E are all located on the last layer k. We assume furthermore that all effect states have the same labels.

Theorem 5. *Let $\mathscr{T} = (S, s_{init}, \to, L)$ be a transition system in which all maximal paths have the same length k. Let E be a set of terminal states and let $C \subseteq S$ be a set of states disjoint from E. Let $\Phi = \Diamond E$ or $\Phi = \Box\neg E$. Let $\pi = s_0 \dots s_n$ be an execution reaching C and satisfying Φ. It is decidable in polynomial time whether C is a d_{Hamm}-counterfactual cause for Φ on π.*

Proof sketch. We sketch the proof for the case that $\Phi = \Diamond E$. We equip the states in S with a weight function $wgt\colon S \to \{0,1\}$ such that the d_{Hamm}-distance of a path to π is equal to the accumulated weight of that path. A state t on layer i gets weight 1 if its label is different to $L(s_i)$. Otherwise, it gets weight 0. Now, we can check whether C is a d_{Hamm}-counterfactual cause, as follows: We remove all states in C and compute a shortest (i.e., weight-minimal) path ζ to E and a shortest path ξ to any terminal state. If the weight of ξ is lower than the weight of ζ, the paths avoiding C that are d_{Hamm}-closest to π do not reach E and C is a d_{Hamm}-counterfactual cause for $\Diamond E$ on π; otherwise, it is not. □

Remark 6. The Hamming distance between paths could easily be extended to account for different levels of similarities between labels: Given a similarity metric d on the set of labels, one could define the distance between two paths $\pi = s_1 \dots s_k$ and $\rho = t_1 \dots t_k$ as $d'_{Hamm}(\pi,\rho) \stackrel{\text{def}}{=} \sum_{i=1}^{k} d(s_i,t_i)$. The algorithm in the proof of Theorem 5 can now easily be adapted to this modified Hamming distance by defining the weight function on the transition system in the obvious way.

Generalized Hamming distance. The assumption in the previous section that all paths in a transition system have the same length is quite restrictive. Hence, we now consider the following generalized version d_{gHamm} of the Hamming distance: For words $w = w_1 \dots w_n$ and $v = v_1 \dots v_m$, we define

$$d_{gHamm}(w,v) \stackrel{\text{def}}{=} \begin{cases} d_{Hamm}(w, v_{[1:n]}) + (m-n) & \text{if } n \leq m, \\ d_{Hamm}(w_{[1:m]}, v) + (n-m) & \text{otherwise.} \end{cases}$$

So d_{gHamm} takes a prefix of the longer word of the same length as the shorter word, computes the Hamming distance of the prefix and the shorter word, and adds the difference in length of the two words.

Theorem 7. *Let $\mathscr{T} = (S, s_{init}, \to, L)$ be a transition system, E a set of terminal states, and C a set of states disjoint from E. Let $\Phi = \Diamond E$ or $\Phi = \Box\neg E$. Let $\pi = s_0 \dots s_n$ be an execution reaching C and satisfying Φ. It is decidable in polynomial time whether C is a d_{gHamm}-counterfactual cause for Φ on π.*

Proof sketch. We adapt the proof of Theorem 5: We take $|\pi|$-many copies of the state space S an let transitions lead from one copy to the next. In the ith copy states with the same label as s_i get weight 0 and all other states get weight 1. Furthermore, we add transitions with weight $|\pi| - i$ from terminal states in a copy $i < |\pi|$ to the same state in the last copy to account for path that are shorter than π. The weight $|\pi| - i$ corresponds to the value added in the generalized Hamming distance when paths of different length are compared. To account for paths longer than π, we furthermore allow transitions with weight 1 within the last copy. These transitions are then taken until a terminal state is reached. With these adaptations, the proof can be carried out analogously to the proof of Theorem 5. □

Levenshtein distance. The idea to check d_{Lev}-counterfactual causality is to construct a weighted transition system to check causality via the computation of shortest paths as for the Hamming distance. So, let $\mathscr{T} = (S, \to, s_{init}, L)$ be a transition system labeled by L with symbols from $\Sigma = 2^{\mathsf{AP}}$. Let E be a set of terminal states and C a set of states disjoint from E. Let $\Phi = \Diamond E$ or $\Phi = \Box\neg E$. Let $\pi = s_1 \dots s_n$ be a maximal path reaching C and satisfying Φ. The transition system we construct contains transitions corresponding directly to the edit operations insertion, deletion and substitution. A path in the constructed transition system then corresponds to an edit sequence between the trace of π and the trace of another path in $\mathscr{T}$. This construction shares some similarities with the construction of Levenshtein automata [36] that accept all words with a Levenshtein distance below a given constant c from a fixed word w.

Now, we formally construct the new weighted transition system $\mathscr{T}^{\pi}_{d_{Lev}}$: The state space of this transition system is $S \times \{1, \dots, n\}$ with the initial state $(s_{init}, 1)$. The labeling function is not used. In $\mathscr{T}^{\pi}_{d_{Lev}}$, we allow the following transitions labelled with letters from the edit alphabet Γ:

1. a transition from (s,i) to $(t,i+1)$ labeled with $(L(s_{i+1}), L(t))$ for each $(s,t) \in \to$ and $i < n$,
2. a transition from (s,i) to (t,i) labeled with $(\varepsilon, L(t))$ for each $(s,t) \in \to$ and $i \leq n$,
3. a transition from (s,i) to $(s,i+1)$ labeled with $(L(s_{i+1}), \varepsilon)$ for each $s \in S$ and $i < n$.

Note that the terminal states in $\mathscr{T}^{\pi}_{d_{Lev}}$ are all contained in $S \times \{n\}$. Any maximal path in $\mathscr{T}^{\pi}_{d_{Lev}}$ corresponds to a maximal path ρ in $\mathscr{T}$. This path ρ is obtained by moving from a state s to a state t in $\mathscr{T}$ whenever a

corresponding transition of type 1 or 2 is taken in $\mathscr{T}^{\pi}_{d_{Lev}}$. Transitions of type 3 do not correspond to a step in $\mathscr{T}$ and stay in the same state.

Furthermore, given a finite path τ in $\mathscr{T}^{\pi}_{d_{Lev}}$ and the corresponding path $\rho = t_1 \dots t_k$ in $\mathscr{T}$, the labels of the transitions of τ form an edit sequence for the words $L(s_2)\dots L(s_n)$ and $L(t_2)\dots L(t_k)$. To see this, observe that, for each $i > 1$, whenever the copy $S \times \{i\}$ is entered in $\mathscr{T}^{\pi}_{d_{Lev}}$, the label of the transition contains $L(s_i)$ in the first component; if a transition stays in a copy $S \times \{i\}$, the label contains ε in the first component. So, the projection onto the first component of the labels of the transitions of τ is indeed $L(s_2)\dots L(s_n)$, potentially with εs in between. In the second component, whenever a transition of type 1 or 2 is taken, the label is simply the label of the corresponding state in ρ. Transitions of type 3 have ε in the second component of their label. Note here that ρ and π both start in s_{init} and that we could hence add $(L(s_{init}), L(s_{init}))$ to the beginning of the edit sequence to obtain an edit sequence for the full traces of τ and ρ. Note that also for infinite paths $\tau = t_1 t_2 \dots$ in $\mathscr{T}^{\pi}_{d_{Lev}}$ the transition labels provide an edit sequence for the words $L(s_2)\dots L(s_n)$ and $L(t_2)L(t_3)\dots$. Vice versa, a finite maximal path $\rho = t_1 \dots t_k$ in $\mathscr{T}$ together with an edit sequence γ for $L(s_2)\dots L(s_n)$ and $L(t_2)\dots L(t_k)$ provides a maximal path τ in $\mathscr{T}^{\pi}_{d_{Lev}}$: The occurrences of ε in γ dictate which type of transition to take while the path ρ tells us which state to move to. As γ projected to the first component contains $L(s_2)\dots L(s_n)$ enriched with εs exactly $n-1$ transitions of type 1 or 3 are taken in τ obtained in this way and we indeed reach the last copy $\mathscr{T} \times \{n\}$. As ρ ends in a terminal state t_k, we furthermore reach the terminal state (t_k, n). Analogously, an infinite path ρ in $\mathscr{T}$ together with an edit sequence γ for $L(\pi)$ and $L(\rho)$ yields an infinite path in $\mathscr{T}^{\pi}_{d_{Lev}}$.

Based on these observations, we equip $\mathscr{T}^{\pi}_{d_{Lev}}$ with a weight function *wgt* on transitions: Transitions labeled with (σ, σ) for a $\sigma \in \Sigma$ get weight 0, the remaining transitions get weight 1.

Theorem 8. *Let $\mathscr{T} = (S, s_{init}, \to, L)$ be a transition system, E a set of terminal states, and C a set of states disjoint from E. Let $\Phi = \Diamond E$ or $\Phi = \Box\neg E$. Let $\pi = s_0 \dots s_n$ be an execution reaching C and satisfying Φ. It is decidable in polynomial time whether C is a d_{Lev}-counterfactual cause for Φ on π.*

Proof sketch. With the construction of the weighted transition system $\mathscr{T}^{\pi}_{d_{Lev}}$ above, the check can be done via the computation of shortest paths as for the Hamming distance above. □

3.3 Relation to Halpern-Pearl causality

In the sequel, we want to demonstrate how our definition of counterfactual causality relates to Halpern-Pearl-style definitions of causality in *structural equation models* [18, 19, 16]. A structural equation model consists of variables $X_1, \dots, X_n$ with finite domains that are governed by equations $X_i = f_i(X_1, \dots, X_{i-1}, C)$ for all $i \leq n$. Here, f_i is an arbitrary function for each i and C is an input parameter for the context. For our consideration, the context C does not play a role and we will hence omit it in the sequel. So, the value of variable X_i depends on the value of (some of) the variables with lower index and the dependency is captured by the function f_i. Halpern and Pearl use *interventions* to define causality for an effect E, which is a set of valuations of $X_1, \dots, X_n$. An intervention puts the value of a variable X_i to some α that is different from $f_i(X_1, \dots, X_{i-1})$, i.e., disregarding the equation f_i. Afterwards, the subsequent variables are evaluated as usual or by further interventions. Halpern and Pearl define:

Definition 9. Let $f_1, \dots, f_n$ over variables $X_1, \dots, X_n$ be a structural equation model as above and let E be an effect set of valuations such that the valuation of $X_1, \dots, X_n$ obtained by the structural equation model belongs to E. A *but-for-cause* is a minimal subset $X \subseteq \{X_1, \dots, X_n\}$ with the following property: There are values α_x for $x \in X$ such that putting variables $x \in X$ to α_x by intervention leads to a valuation of $X_1, \dots, X_n$ not exhibiting the effect E. More precisely, letting t_i be the valuation $[X_1 = w_1, \dots, X_{i-1} = w_{i-1}]$, where $w_i = f_i(w_1, \dots, w_{i-1})$ if $X_i \notin X$, and $w_i = \alpha_{X_i}$ if $X_i \in X$, we get that $t_{n+1} \notin E$.

In order to compare this to our notion of counterfactual causes, we view structural equation models as tree-like transition system $\mathscr{T}$: The nodes at level i are valuations for the variables $X_1,\dots,X_{i-1}$. At each node s at level i, two actions are available: The action default moves to the state on level $i+1$ where the valuation in s is extended by setting X_i to the value $f_i(X_1,\dots,X_{i-1})$ where the values for $X_1,\dots,X_{i-1}$ are taken from the valuation in s. The action intervention extends the valuation of s by setting X_i to any other value than the action default. The labelling in $\mathscr{T}$ assigns the label $\{\text{intervention}\}$ to all states that are reached by the action intervention. The remaining states and the initial state with the empty valuation get the label $\emptyset$. Given an effect E as a set of valuations, we interpret this as the corresponding set of leaf states in $\mathscr{T}$. The default path π that always chooses the action default corresponds to evaluating the equations in the structural equation model without interventions. We can now capture but-for-causality with d_{Hamm}-counterfactual causality along the default path π if all variables are Boolean:

Proposition 10. *Let $f_1,\dots,f_n$ over Boolean variables $X_1,\dots,X_n$ be a structural equation model and let E be an effect set of valuations. Let X be a but-for-cause for E. Let C_X be the set of all nodes in the transition system $\mathscr{T}$ which are reached by a* default*-transition for a variable $x \in X$. Then, C_X is a d_{Hamm}-counterfactual cause for E in $\mathscr{T}$ on the default path π.*

For non-Boolean variables, the definitions of but-for-causes and of d_{Hamm}-counterfactual causes have one significant difference: A but-for-cause X merely requires the existence of values to assign to the variables in X by intervention such that the effect is avoided. A d_{Hamm}-counterfactual cause C in $\mathscr{T}$ requires that for all possible interventions on the variables in X, the effect is avoided. This universal quantification originates from the universal quantification over most similar worlds in the Stalnaker-Lewis semantics of counterfactual causality.

The minimality requirement of but-for-causes does not have a counterpart in the definition of d-counterfactual causes. This allows us to assert that a candidate set of states C is a d-counterfactual cause for an effect even if it contains redundancies. When trying to find d-counterfactual causes for a given effect, on the other hand, of course trying to find (cardinality-)minimal causes is a reasonable option.

Besides but-for causality, we can also capture actual causality as in [16] in our framework in the case of Boolean structural equation models. This is demonstrated in the extended version [31].

4 Counterfactual causality in reachability games

The counterfactual notion of causality introduced and investigated in the previous section can be applied to reachability games $\mathscr{G}$: We take the perspective of a player Π. Given a strategy σ for the opponent and a play in which Π lost, we apply the definition to the transition system obtained from σ and $\mathscr{G}$ and the given play. This allows us to analyze whether avoiding a certain set of states while playing against strategy σ as similarly as possible to the given play would have allowed Π to win. Depending on whether we take the perspective of Reach or Safe, the effect that the player loses the game is a safety or reachability property, which we considered as effects in transition systems. The need to be given a strategy for the opponent, however, constitutes a major restriction to the usefulness of this approach. All proofs omitted in this section can be found in the extended version [31].

4.1 D-counterfactual causality

We provide a definition of counterfactual causality in reachability games in the sequel in which we only need the strategy σ with which the player Π played and are interested in why the strategy σ allows the

opponent to win the game. Since both players have optimal MD-strategies in a reachability game, we restrict ourselves to MD-strategies in the definition.

Definition 11. Let $\mathscr{G}$ be a reachability game with target set V_{Eff}. Let Π be one of the two players and let σ be a MD-strategy for player Π. Let C be a set of locations disjoint from V_{Eff}. Let D be a distance function on MD-strategies. We say that C is a D-counterfactual cause for the fact that Π loses using σ if

1. there are σ-plays that reach C on which Π loses,
2. there is an MD-strategy τ for player Π that avoids C (i.e., there is no τ-play reaching C),
3. all MD-strategies τ for player Π, that avoid C and that have minimal D-distance to σ among the strategies avoiding C, are winning for Π.

If we take the perspective of player Π in game $\mathscr{G}$ where the opponent $\bar{\Pi}$ does not control any locations, MD-strategies for Π satisfying condition 1 of the definition are essentially simple paths satisfying a safety or reachability effect property (with additional information on the states that are not visited by the path). To some extent, the definition can now be seen as a generalization of the definition for transition systems for suitable distance functions D: We say a strategy distance function D *generalizes a path distance function* d if in games where $\bar{\Pi}$ does not control any location, for all strategies σ, τ for Π, we have $D(\sigma,\tau) = d(\pi_\sigma, \pi_\tau)$ where π_σ and π_τ are the unique σ- and τ-plays. The definition that C is a D-counterfactual cause for σ losing the game agrees with the definition that C is a d-counterfactual cause on π_σ for $\Diamond V_{Eff}$ or $\Box\neg V_{Eff}$ in acyclic games in this case. In cyclic games, there is one caveat: The definition for games quantifies only over MD-strategies which induce a play that is a simple path or simple lasso. The definition for transition systems quantifies over more complicated paths as well.

Hausdorff distance d^H_{pref} based on the prefix metric d_{pref}. A way to obtain a strategy distance function generalizing a given path distance function is the use of the Hausdorff distance on the set of plays of the strategies [12, Section 6.2.2]: Let τ and σ be two MD-strategies, and d be a distance function over plays. The Hausdorff distance d^H based on d is defined by

$$d^H(\sigma,\tau) = \max\left\{ \sup_{\sigma\text{-plays } \pi} \inf_{\tau\text{-plays } \rho} d(\pi,\rho),\ \sup_{\tau\text{-plays } \rho} \inf_{\sigma\text{-plays } \pi} d(\pi,\rho) \right\}.$$

Let us consider the Hausdorff distance d^H_{pref} based on the prefix metric d_{pref} assuming that all states have a unique label. For two strategies σ and τ for Safe, the distance $d^H_{pref}(\sigma,\tau)$ is 2^{-n} where n is the least natural number such that there is a prefix of length n of a τ-play that is not a prefix of a σ-play, or vice versa. In order to find strategies that are as similar as possible to a given strategy σ, we hence have to consider strategies that follow σ for as many steps as possible. This leads to an algorithm for checking d^H_{pref}-counterfactual causality in reachability games that shares some similarities with the algorithm for checking d^{AP}_{pref}-counterfactual causality in transition systems.

Theorem 12. *Let $\mathscr{G} = (V, v_i, \Delta)$ where $V = V_{\mathsf{Reach}} \uplus V_{\mathsf{Safe}} \uplus V_{Eff}$ be a reachability game with target set V_{Eff} and σ a MD-strategy for player Π. Let C be a set of locations disjoint from V_{Eff}. We can check in polynomial time whether C is a d^H_{pref}-counterfactual cause for the fact that Π is losing using σ in $\mathscr{G}$.*

For the Hausdorff lifting of d_{Hamm} or d_{Lev}, the resulting notion of counterfactual causes in games is more complicated. If we try to adapt the approach used in transition systems, we need a way to capture the minimum distance of a given strategy to the closest winning strategies. However, shortest path games (as extension of the weighted transition systems used for d_{Hamm}- and d_{Lev}-counterfactual causes in transition systems) cannot be employed in an obvious way. In this paper, we now instead consider two further distance functions related to the Hamming distance for which we can provide algorithmic results.

Hamming strategy distance. Let σ and τ be two MD-strategies for Π in $\mathscr{G}$, we define the Hamming strategy distance by $d^s_{Hamm}(\sigma,\tau) = |\{v \in V \mid \sigma(v) \neq \tau(v)\}|$. As the Hamming distance on paths counts positions at which traces differ, the Hamming strategy distance counts positions at which two MD-strategies differ. Using a similar proof using shortest-path games [23] as for Theorem 5, we obtain the following polynomial-time result in the case of *aperiodic* games.

Theorem 13. *Let $\mathscr{G} = (V, v_i, \Delta)$ be an acyclic reachability game with target set V_{Eff} and σ a MD-strategy for player Π. Let C be a set of locations disjoint from V_{Eff}. We can check in polynomial time whether C is a d^s_{Hamm}-counterfactual cause for the fact that Π is losing using σ in $\mathscr{G}$.*

Hausdorff-inspired distance d^*. The distance function d^* computes the number of vertices where two MD-strategies make distinct choices along each play of both MD-strategies. It hence has some similarity to a Hausdorff-lifting of the Hamming distance on paths. This Hausdorff-lifting, however, counts the number of *occurrences* of vertices at which two paths differ (in their label). Instead, for a play $\rho = v_0 v_1 \dots$ and a strategy σ for Π, we define the *distance between σ and ρ $dist(\rho,\sigma)$* as the number of vertices $v \in V_\Pi$ (i.e., not the number of occurrences) such that there exists $i \in \mathbb{N}$ with $v = v_i$ in ρ, and $\sigma(v_i) \neq (v_i, v_{i+1})$.We define d^* for two strategies τ, σ by

$$d^*(\tau,\sigma) = \max\big(\sup_{\rho|\tau\text{-play}} dist(\rho,\sigma), \sup_{\pi|\sigma\text{-play}} dist(\pi,\tau)\big).$$

To simplify the notation, we define $d^\tau(\sigma) = \sup_{\rho|\tau\text{-play}} dist(\rho,\sigma)$. We prove that the threshold problem for d^* is NP-complete via a reduction from the *longest simple path problem*:

Proposition 14. *Let $\mathscr{G}$ be a reachability game, σ, τ be two MD-strategies for Π, and $k \in \mathbb{N}$ be a threshold. Then deciding if $d^*(\tau,\sigma) \geq k$ is NP-complete.*

The proposition explains why understanding d^*-counterfactual causes is complex. We leave a further investigation of such notions for future work. As a first step toward a better understanding, we turn our attention to a conceptually simpler notion, the explanation induced by a counterfactual cause.

Example 15. Let us illustrate counterfactual causes according to distances on strategies. We consider the reachability game depicted in the left of Figure 2 and the non-winning strategy σ for Reach depicted in green. Under d^H_{pref} or d^*, the counterfactual cause for Reach is $\{v_2, v_3\}$. Indeed, there exists one play that reaches v_3 and loses for Reach, and there exists a unique strategy that avoids $\{v_2, v_3\}$ by changing the choice of σ in v_1. Moreover, this counterfactual cause is minimal since $\{v_3\}$ is not a cause. Indeed, the (losing) strategy that differs from σ in v_0 and v_1 avoids $\{v_3\}$ with a minimal distance to σ, i.e. 2^{-2} for d^H_{pref} and 1 for d^*. Under d^s_{Hamm}, the counterfactual cause for Reach is $\{v_3\}$. Indeed, two strategies exist with a distance of 1 to σ according to the vertex where Reach changes its decision. In these two strategies, only one avoids $\{v_3\}$: the strategy where Reach change its decision in v_1. ⌟

4.2 D-counterfactual explanation

Given a D-counterfactual cause, we want to explain what is wrong in the losing strategy for Π. In particular, we are interested in sets of locations C such that Π could have won the game if she had not made the decisions of σ in the locations in C.

Definition 16. Let $\mathscr{G}$ be a reachability game and σ be a non-winning MD-strategy for Π. Let $E \subseteq V_\Pi$. We call E an *explanation* in $\mathscr{G}$ under σ if there exists a winning MD-strategy τ such that for all vertices $v \in V_\Pi$, $\tau(v) = \sigma(v)$ iff $v \notin E$. We call such a τ an *E-distinct σ-strategy*.

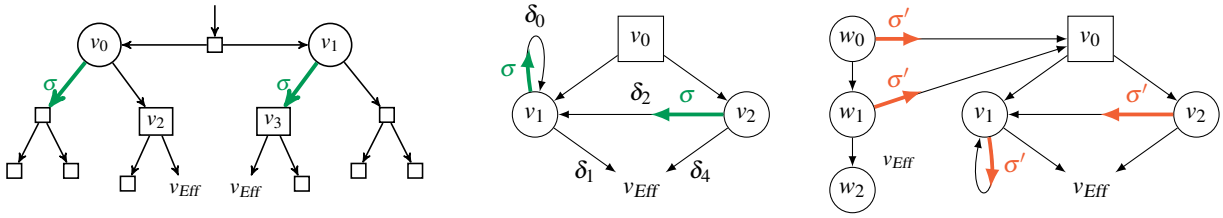

Figure 2: On the left and in the middle, two reachability games with initial vertex v_0 and strategy σ for Reach (depicted in green). On the right, the reachability game obtained by reduction of Corollary 21 from the game depicted in the middle with initial vertex w_0 and σ' be a non winning strategy for Reach.

We note that the definition of an explanation does not refer to a distance function. However, given a D-counterfactual cause, we can compute an explanation no matter which distance D is used.

Proposition 17. *Let $\mathcal{G} = (V, v_i, \Delta)$ be a reachability game, D a distance function on strategies and σ be a non-winning MD-strategy for Π. Let $C \subseteq V$ be a D-counterfactual cause. We can compute an explanation E (from C) in polynomial time.*

Proof. Let $\mathcal{G}' = (V \setminus C, v_i, \Delta)$ be the reachability game. Since D is a D-counterfactual cause, we know that there exists a winning strategy τ in $\mathcal{G}'$. We can compute this strategy in time polynomial in the size of $\mathcal{G}'$ with the attractor method and we define $E = \{v \mid \sigma(v) \neq \tau(v)\}$. □

A winning strategy differing from σ in E might not have much in common with σ. For this reason, explanations that point out changes in the decisions of σ in E that enforce only the minimal necessary change to obtain a winning strategy τ from σ are of particular interest. We can use a distance function D to quantify how much a strategy needs to be changed.

Definition 18. Let $\mathcal{G}$ be a reachability game and σ be a non-winning MD-strategy for Π. For a distance function D for MD-strategies, we call a explanation E a *D-minimal explanation*, if there exists a winning E-distinct σ-strategy τ with $d(\tau, \sigma) = \min\{d(\mu, \sigma) \mid \mu \text{ is a winning MD-strategy for } \Pi\}$.

For a strategy σ and an explanation E, the distance $d^s_{Hamm}(\sigma, \tau)$ for an E-distinct σ-strategy τ is precisely $|E|$. So, d^s_{Hamm}-minimal explanations are cardinality-minimal explanations.

Example 19. Let us illustrate explanations and D-minimal explanations. We consider the reachability game $\mathcal{G}$ where Reach wins depicted in the left of Figure 2 with σ, a non-winning MD-strategy for Reach, depicted in green. We note that $E = \{v_1, v_2\}$ is an explanation in $\mathcal{G}$ under σ. A winning E-distinct σ-strategy τ for Reach is given by $\tau(v_1) = \delta_1$ and $\tau(v_2) = \delta_4$. However, E is not a d^*-minimal explanation or d^s_{Hamm}-minimal explanation. Clearly, $d^s_{Hamm}(\tau, \sigma) = 2$. Further, also $d^*(\tau, \sigma) = 2$ as the σ-play $v_0 v_2 v_1^\omega$ visits two states, namely v_2 and v_1 at which σ and τ make different decisions. The set $E' = \{v_1\}$, however is a d^*-minimal explanation and d^s_{Hamm}-minimal explanation: the E'-distinct σ-strategy τ' choosing δ_1 in v_1 and behaving like σ in v_2 wins and has d^s_{Hamm}- and d^*-distance 1 to σ. As any winning strategy has at least distance 1 to σ, E' is hence a D-minimal explanation for both distance functions. ⌟

For D-minimal explanations, it is central to find a winning MD-strategy that minimises the distance D to the given losing strategy σ. We take a look at this problem from the point of view of Reach and prove that for d^s_{Hamm} and d^* the associated threshold problems are not in P if P≠NP.

Theorem 20. *Given a game $\mathcal{G}$, a losing strategy σ for* Reach*, and $k \in \mathbb{N}$, deciding if there exists a winning MD-strategy τ for* Reach *such that $d^s_{Hamm}(\tau, \sigma) \leq k$ is NP-complete. Further, the problem whether there is a winning MD-strategy τ with $d^*(\tau, \sigma) \leq k$ is not in P if P≠NP.*

Proof sketch. To establish the NP upper bound for d^s_{Hamm}, we can guess a MD-strategy τ for Reach and check in polynomial time whether it is winning and whether $d^s_{Hamm}(\tau,\sigma) \leq k$. For the NP-hardness for d^s_{Hamm}, we provide a polynomial-time many-one reduction from the NP-complete decision version of the *feedback vertex set* [22]. Given a cyclic (directed) graph G, this problem asks whether there is a set S of size at most k such that if we remove this set, $G \setminus S$ becomes acyclic. For the problem for d^*, we provide a polynomial-time Turing reduction from the same problem. A detailed proof is given in [31]. □

We deduce that checking D-minimality of an explanation cannot be done in polynomial time if P≠NP.

Corollary 21. *Let $\mathcal{G}$ be a reachability game, σ be a non-winning MD-strategy for* Reach*, and $E \subseteq V$. The problem to check if E is a d^s_{Hamm}-minimal explanation in $\mathcal{G}$ for σ is coNP-complete. The problem to check if E is a d^*-minimal explanation in $\mathcal{G}$ for σ is not in P if P≠NP.*

Despite the hardness in the general case, if $\mathcal{G}^\sigma$ is acyclic, we prove that we can compute the winning MD-strategy that minimises the d^*-distance to σ in polynomial time. From this strategy, a d^*-minimal explanation can then be computed as in Proposition 17. The proof of Theorem 22 (in the extended version [31]) constructs a shortest-path game [23] without negative weights in which an optimal strategy, that leads to the desired winning strategy in the original game, can be computed in polynomial time.

Theorem 22. *Let $\mathcal{G}$ be reachability game where* Reach *wins, and σ be a non-winning MD-strategy for* Reach *such that $\mathcal{G}^\sigma$ is acyclic. Then, we can compute a winning MD-strategy τ that minimizes the distance d^* to σ in polynomial time.*

5 Conclusion and Outlook

The notion of d-counterfactual cause for a distance function d in transition systems turned out to be checkable in polynomial time for the distance functions d_{pref}, d_{Hamm}, and d_{Lev} and so it has the potential to be employed in efficient tools to provide understandable explanations of the behavior of a system. In our algorithmic results for safety effects Φ, one caveat remains: we only considered finite executions reaching a cause candidate C and satisfying Φ. Allowing also finitely representable, e.g., ultimately periodic paths, constitutes a natural extension, which requires adjustments in the provided algorithms.

The problem of finding good causes remains as future work: Whenever causality can be checked in polynomial time, there is an obvious non-deterministic polynomial-time upper bound on the problem to decide whether there are causes below a given size, but the precise complexities are unclear. A further idea is to use the distance function to assess how good a cause is by considering the distance from the actual execution to the closest executions avoiding a cause. For reachability effects and the prefix and Hamming distance, the set of direct predecessors optimizes this distance. For other distance functions or safety causes, this measure could, nevertheless, be more useful. The search for similar measures for the quality of causes constitutes an interesting direction for future work.

In reachability games, we saw that the analogous definition of D-counterfactual causes can be checked in polynomial time for the Hausdorff-lifting d^H_{pref} of the prefix metric, as well. For other distance functions, the definition seems to lead to complicated notions due to the involved quantification over all MD-strategies avoiding the cause and having a minimal distance to a given strategy. A closer investigation of these notions might, nevertheless, be a fruitful subject for future research. However, our analysis of the conceptually simpler D-minimal explanations provides insights into the complications one might encounter here. For the Hausdorff-inspired distance function d^*, we showed that already the threshold problem for the distance between two given MD-strategies is NP-hard. Furthermore, for the relatively simple distance function d^s_{Hamm}, checking the d^s_{Hamm}-minimality of an explanation is in coNP-complete. For the Hausdorff-inspired distance function d^*, checking d^*-minimality is not in P unless P=NP.

References

[1] Christel Baier, Norine Coenen, Bernd Finkbeiner, Florian Funke, Simon Jantsch & Julian Siber (2021): *Causality-based game solving*. In: *International Conference on Computer Aided Verification*, Springer, pp. 894–917, doi:10.1007/978-3-030-81685-8_42.

[2] Christel Baier, Clemens Dubslaff, Florian Funke, Simon Jantsch, Rupak Majumdar, Jakob Piribauer & Robin Ziemek (2021): *From Verification to Causality-Based Explications (Invited Talk)*. In Nikhil Bansal, Emanuela Merelli & James Worrell, editors: *48th International Colloquium on Automata, Languages, and Programming, (ICALP)*, *LIPIcs* 198, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 1:1–1:20. Available at `https://doi.org/10.4230/LIPIcs.ICALP.2021.1`.

[3] Christel Baier, Clemens Dubslaff, Florian Funke, Simon Jantsch, Jakob Piribauer & Robin Ziemek (2022): *Operational Causality – Necessarily Sufficient and Sufficiently Necessary*. In Nils Jansen, Mariëlle Stoelinga & Petra van den Bos, editors: *A Journey from Process Algebra via Timed Automata to Model Learning : Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday*, Springer Nature Switzerland, Cham, pp. 27–45, doi:10.1007/978-3-031-15629-8_2.

[4] Christel Baier, Florian Funke & Rupak Majumdar (2021): *A Game-Theoretic Account of Responsibility Allocation*. In Zhi-Hua Zhou, editor: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, International Joint Conferences on Artificial Intelligence Organization, pp. 1773–1779, doi:10.24963/ijcai.2021/244. Main Track.

[5] Christel Baier & Joost-Pieter Katoen (2008): *Principles of Model Checking*. MIT Press.

[6] Thomas Ball, Mayur Naik & Sriram K. Rajamani (2003): *From Symptom to Cause: Localizing Errors in Counterexample Traces*. *SIGPLAN Not.* 38(1), pp. 97–105, doi:10.1145/640128.604140.

[7] Ilan Beer, Shoham Ben-David, Hana Chockler, Avigail Orni & Richard J. Trefler (2012): *Explaining counterexamples using causality*. *Formal Methods in System Design* 40(1), pp. 20–40, doi:10.1007/s10703-011-0132-2.

[8] Hana Chockler (2016): *Causality and Responsibility for Formal Verification and Beyond*. In Gregor Gößler & Oleg Sokolsky, editors: *Proceedings First Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies, CREST@ETAPS 2016, Eindhoven, The Netherlands, 8th April 2016*, *EPTCS* 224, pp. 1–8, doi:10.4204/EPTCS.224.1.

[9] E. M. Clarke, O. Grumberg & D. Peled (1999): *Model Checking*. MIT Press.

[10] Edmund M. Clarke, Orna Grumberg, Kenneth L. McMillan & Xudong Zhao (1995): *Efficient Generation of Counterexamples and Witnesses in Symbolic Model Checking*. In: *Proc. of the 32nd Annual ACM/IEEE Design Automation Conf. (DAC)*, ACM, New York, NY, USA, pp. 427–432, doi:10.1145/217474.217565.

[11] Norine Coenen, Bernd Finkbeiner, Hadar Frenkel, Christopher Hahn, Niklas Metzger & Julian Siber (2022): *Temporal Causality in Reactive Systems*. In: *20th International Symposium on Automated Technology for Verification and Analysis, ATVA*, pp. 25–28, doi:10.1007/978-3-031-19992-9_13.

[12] M.C. Delfour & J.P. Zolesio (2011): *Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization, Second Edition*. Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), doi:10.1137/1.9780898719826. Available at `https://books.google.fr/books?id=fjjvX9a9cxUC`.

[13] Erich Grädel, Wolfgang Thomas & Thomas Wilke, editors (2002): *Automata Logics, and Infinite Games: A Guide to Current Research*. Springer-Verlag, Berlin, Heidelberg.

[14] Alex Groce, Sagar Chaki, Daniel Kroening & Ofer Strichman (2006): *Error explanation with distance metrics*. *International Journal on Software Tools for Technology Transfer* 8(3), pp. 229–247, doi:10.1007/978-3-540-24730-2_8.

[15] Alex Groce & Willem Visser (2003): *What Went Wrong: Explaining Counterexamples*. In Thomas Ball & Sriram K. Rajamani, editors: *Model Checking Software*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 121–136, doi:10.1007/3-540-44829-2_8.

[16] Joseph Y. Halpern (2015): *A Modification of the Halpern-Pearl Definition of Causality*. In: *Proc. of the 24th Intern. Joint Conf. on AI (IJCAI)*, AAAI Press, pp. 3022–3033.

[17] Joseph Y. Halpern & Judea Pearl (2001): *Causes and Explanations: A Structural-Model Approach: Part i: Causes*. In: *Proc. of the 17th Conf. on Uncertainty in AI (UAI)*, Morgan Kaufmann Publishers Inc., pp. 194–202, doi:10.1093/bjps/axi147.

[18] Joseph Y. Halpern & Judea Pearl (2005): *Causes and Explanations: A Structural-Model Approach. Part I: Causes*. *The British Journal for the Philosophy of Science* 56(4), pp. 843–887, doi:10.1093/bjps/axi147.

[19] Joseph Y. Halpern & Judea Pearl (2005): *Causes and Explanations: A Structural-Model Approach. Part II: Explanations*. *The British Journal for the Philosophy of Science* 56(4), pp. 889–911, doi:10.1093/bjps/axi148.

[20] David Hume (1739): *A Treatise of Human Nature*. John Noon, doi:10.1093/oseo/instance.00032872.

[21] David Hume (1748): *An Enquiry Concerning Human Understanding*. London.

[22] Richard M. Karp (1972): *Reducibility among Combinatorial Problems*, pp. 85–103. Springer US, Boston, MA, doi:10.1007/978-1-4684-2001-2_9.

[23] Leonid Khachiyan, Endre Boros, Konrad Borys, Khaled Elbassioni, Vladimir Gurvich, Gabor Rudolf & Jihui Zhao (2007): *On Short Paths Interdiction Problems: Total and Node-Wise Limited Interdiction*. *Theory of Computing Systems*, doi:10.1007/s00224-007-9090-x.

[24] Florian Leitner-Fischer & Stefan Leue (2013): *Causality Checking for Complex System Models*. In: *Proc. of the 14th Intern. Conf. on Verification, Model Checking, and Abstract Interpretation (VMCAI)*, pp. 248–267, doi:10.1007/978-3-642-35873-9_16.

[25] Vladimir I. Levenshtein (1966): *Binary codes capable of correcting deletions, insertions, and reversals*. *Soviet physics doklady* 10(8), pp. 707–710.

[26] David Lewis (1973): *Causation*. *Journal of Philosophy* 70(17), pp. 556–567, doi:10.2307/2025310.

[27] David K. Lewis (1973): *Counterfactuals*. Cambridge, MA, USA: Blackwell.

[28] J. L. Mackie (1965): *Causes and Conditions*. *American Philosophical Quarterly* 2(4), pp. 245–264. Available at `http://www.jstor.org/stable/20009173`.

[29] Z. Manna & A. Pnueli (1995): *The Temporal Logic of Reactive and Concurrent Systems: Safety*. Springer-Verlag.

[30] Kedar S. Namjoshi (2001): *Certifying Model Checkers*. In: *13th International Conference on Computer Aided Verification (CAV)*, *Lecture Notes in Computer Science* 2102, Springer, pp. 2–13. Available at `https://doi.org/10.1007/3-540-44585-4_2`.

[31] Julie Parreaux, Jakob Piribauer & Christel Baier (2023): *Counterfactual Causality for Reachability and Safety based on Distance Functions*. arXiv:2308.11385. ArXiv preprint: `arxiv.org/abs/2308.11385`.

[32] Judea Pearl (2009): *Causality*, 2 edition. Cambridge University Press, doi:10.1017/CBO9780511803161.

[33] Jonas Peters, Dominik Janzing & Bernhard Schölkopf (2017): *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

[34] Ibo van de Poel (2011): *The Relation Between Forward-Looking and Backward-Looking Responsibility*, pp. 37–52. Springer Netherlands, Dordrecht, doi:10.1007/978-94-007-1878-4_3.

[35] Manos Renieres & Steven P. Reiss (2003): *Fault localization with nearest neighbor queries*. In: *Proc. of the 18th IEEE Intern. Conf. on Automated Software Engineering (ASE)*, pp. 30–39, doi:10.1109/ASE.2003.1240292.

[36] Klaus U Schulz & Stoyan Mihov (2002): *Fast string correction with Levenshtein automata*. *International Journal on Document Analysis and Recognition* 5(1), pp. 67–85, doi:10.1007/s10032-002-0082-8.

[37] Robert C. Stalnaker (1968): *A Theory of Conditionals*. In William L. Harper, Robert Stalnaker & Glenn Pearce, editors: *IFS. The University of Western Ontario Series in Philosophy of Science*, 15, Springer, Dordrecht, pp. 41–55, doi:10.1007/978-94-009-9117-0_2.

[38] Chao Wang, Zijiang Yang, Franjo Ivancic & Aarti Gupta (2006): *Whodunit? Causal Analysis for Counterexamples*. In: *Proc. of the 4th Intern. Symp. on Automated Technology for Verification and Analysis (ATVA)*, pp. 82–95, doi:10.1007/11901914_9.

[39] Shaohui Wang, Anaheed Ayoub, BaekGyu Kim, Gregor Gößler, Oleg Sokolsky & Insup Lee (2013): *A Causality Analysis Framework for Component-Based Real-Time Systems*. In: *Proceedings of the 4th International Conference on Runtime Verification (RV)*, pp. 285–303, doi:10.1007/978-3-642-40787-1_17.

[40] Vahid Yazdanpanah & Mehdi Dastani (2016): *Distant Group Responsibility in Multi-agent Systems*. In Matteo Baldoni, Amit K. Chopra, Tran Cao Son, Katsutoshi Hirayama & Paolo Torroni, editors: *PRIMA 2016: Princiles and Practice of Multi-Agent Systems - 19th International Conference, Phuket, Thailand, August 22-26, 2016, Proceedings*, *Lecture Notes in Computer Science* 9862, Springer, pp. 261–278, doi:10.1007/978-3-319-44832-9_16.

[41] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina & Brian Logan (2019): *Strategic Responsibility Under Imperfect Information*. In Edith Elkind, Manuela Veloso, Noa Agmon & Matthew E. Taylor, editors: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 592–600. Available at `http://dl.acm.org/citation.cfm?id=3331745`.

[42] Andreas Zeller (2002): *Isolating Cause-Effect Chains from Computer Programs*. In: *Proc. of the 10th ACM SIGSOFT Symp. on Foundations of Software Engineering (FSE)*, ACM, New York, NY, USA, pp. 1–10, doi:10.1145/587051.587053.