Detecting causal relationships in simulation models using intervention-based counterfactual analysis

BENJAMIN C. HERD and SIMON MILES, King's College London, UK

Central to explanatory simulation models is their capability to not just show *that* but also *why* particular things happen. Explanation is closely related with the detection of causal relationships and is, in a simulation context, typically done by means of controlled experiments. However, for complex simulation models, conventional 'blackbox' experiments may be too coarse-grained to cope with spurious relationships. We present an intervention-based causal analysis methodology that exploits the manipulability of computational models and detects and circumvents spurious effects. The core of the methodology is a formal model that maps basic causal assumptions to causal observations and allows for the identification of combinations of assumptions that have a negative impact on observability. First experiments indicate that the methodology can successfully deal with notoriously tricky situations involving asymmetric and symmetric overdetermination and detect fine-grained causal relationships between events in the simulation. As illustrated in the paper, the methodology can be easily integrated into an existing simulation environment.

CCS Concepts: • **General and reference → Verification**; • **Computing methodologies → Model verification and validation**;

Additional Key Words and Phrases: Causal Analysis, Simulation, Counterfactuals

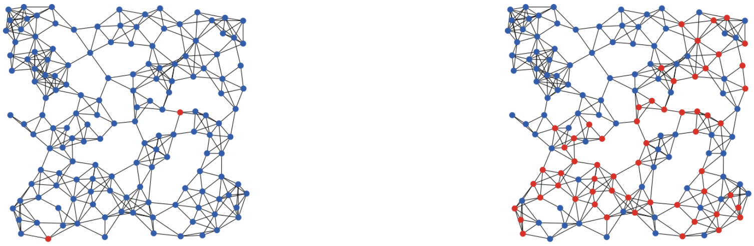
ACM Reference Format:
Benjamin C. Herd and Simon Miles. 2018. Detecting causal relationships in simulation models using intervention-based counterfactual analysis. *ACM Trans. Intell. Syst. Technol.* 1, 1 (April 2018), 28 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

One of the great benefits of cheap modern computer hardware is the possibility to use simulation as a virtual testbed for the exploration of various different scenarios. Computer simulation makes it possible to explore complex real-world scenarios and conduct what-if analyses in an efficient and cost-effective way. This is particularly interesting when real-world experiments cannot be carried out, e.g. because of financial, legal, or ethical constraints. For the simulation of complex adaptive systems that exhibit non-linear and emergent behaviour, *agent-based modelling (ABM)* has emerged as a powerful paradigm [17]. It is a bottom up technique that models the constituents of a complex system as individual agents with their actions and interactions. As a

Authors' address: Benjamin C. Herd, benjamin.c.herd@kcl.ac.uk; Simon Miles, simon.miles@kcl.ac.uk, King's College London, Strand, London, WC2R 2LS, UK.

2157-6904/2018/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Initial state

(b) Final state after 300 time steps

Fig. 1. An agent-based simulation of virus propagation in a network

motivating example, consider an ABM of virus propagation on a network as shown in Figure 1. Each node in the graph represents an agent, each edge represents a direct contact relationship between the two agents. Each agent can be in one of two states: infected or healthy. In each time step, infected agents try to infect their directly connected healthy neighbours which may, in turn, become infected with a certain probability. The transmission of a disease from an infected agent to a healthy thus represents a causal relationship between agents. Infected agents may recover with a certain probability. Let us now look at a particular simulation run. In the beginning of the simulation, three agents are infected (coloured red in Figure 1a), the remaining agents are healthy (coloured blue). The state of the population after running the simulation for 300 time steps is shown in Figure 1b.

By 'growing' the complex emergent system from the individual actions and interactions of the underlying agent population, ABM has great *explanatory* potential. However, despite its increasing adoption in various domains [3, 4, 19, 28], many people are still very sceptical of the benefits of ABM as opposed to more traditional, better understood techniques such as differential equations or econometrics. A main reason for that scepticism is that, due to their high level of complexity, ABMs are notoriously difficult to engineer, to understand, and to control. As a consequence, agent-based models are often seen as black boxes whose internal dynamics are insufficiently well understood [15, 27]. When constructing models, users are often confronted with an overwhelming amount of information that needs to be visualised, processed, and made sense of in order for the model to serve its explanatory purposes. An important aspect of explanation is the detection of *causal relationships*. Causal relationships can be subdivided into cases of *type causation*, i.e. causal connections between *general kinds of events* such as the link between smoking and lung cancer, and cases of *token causation*, i.e. causal connections between *specific instantiated events* such as the link between the toss of a particular stone and the shattering of a particular window at a particular point in time. In this paper, we focus on *token causation*, i.e. *causal relationships between specific events occurring at a particular time during a specific simulation run.*

Returning to the virus propagation example, a modeller may, for example, be interested in explaining the observed transmission dynamics by revealing certain cases of token causation. In order to understand how the disease spreads, she may be interested in determining for each of the eventually infected agents in Figure 1b, which of the initially infected agents in Figure 1a served as the respective 'root cause', i.e. initiated the causal chain that finally led to the agent in question becoming infected. One way to answer this question would be to perform three experiments, each one with a different initially infected agent's health state being manipulated (i.e. switched to 'healthy') in order to investigate the effect of that change on the infection dynamics. However, there are some issues with that approach. First, due to the probabilistic decisions in the simulation model, each of the three experiments would require a sufficiently large number of simulation runs to be conducted, followed by statistical analysis of the output data in order to decide whether the respective null hypothesis ('initially infected agent x was not the root cause') can be rejected. Depending on the size of the model, this may be time-consuming. Second, following this statistical process, a stipulated observed causal relationship can only be *falsified* but not *verified*. A solution could be to run many experiments with varied initial conditions and investigate statistical regularities , i.e. correlations, between initially and eventually infected agents. Apart from this approach being fairly costly, it is well known that correlations may well *indicate* causal relationships but can never strictly *prove* them. As summarised nicely by Guerini and Moneta, "indeed the reproduction, no matter how robust, of a set of statistical properties of the data by a model is a quite weak form of validation, since, in general, given a set of statistical dependencies there are possibly many causal structures which may have generated them." [7] A third problem is that, even if the results point strongly towards a causal relationship between two infection events, that conclusion may still be wrong due to confounding effects such as *preemption*. For example, manipulating the disease transmission model such that one originally infected agent is made healthy in the beginning may open the door for spurious background effects to creep in and, from an observer's perspective, 'mask' the original causal relationship.

So, in summary, investigating causal relationships between specific events in a simulation experiment is exacerbated by two issues that need to be dealt with appropriately: (i) *randomness* which complicates the establishment of strict existence proofs of causal relationship, and (ii) *confounding effects* which 'mask' true causal relationships. The first issue is fairly easy to deal with by fixing the random seed and ensuring that similar streams of pseudo-random numbers are generated in different simulation runs. The main focus of this paper is on the second issue. As for the simulation to be analysed, we make the following assumptions:

(1) The simulation is finite and proceeds in discrete time steps.
(2) The state of the simulation at any time is characterised by a set of binary events, i.e. propositions that are either true or false[1].
(3) The simulation state is observable, i.e. the simulation records and outputs all relevant events at every time step, resulting in a finite simulation trace.
(4) The simulation can be manipulated such that arbitrary events can be prevented from occurring while the simulation is running.

The central research question of this paper can now be stated as follows.

> Let m be a simulation model whose execution produces a trace π of length k which represents a finite sequence of k sets of events, one set

[1] Generalisation to multi-valued events is straightforward. For clarity but without loss of generality, we restrict our focus to binary events.

> per simulated time step. Let further e_t and $e_{t'}$ be two events in π that happened at time steps t and $t' > t$, respectively. Did e_t cause $e_{t'}$ in the particular run represented by π?

To answer this question, we present a **simulation-based causal analysis methodology** that is capable of reliably detecting token-causal relationships between active events in a simulation model. Rather than relying upon statistical analysis, the methodology is based on *automated active intervention* in the simulation process. Intuitively, the simulation model is 'nudged' during its execution such that true causal relationships between active events become detectable reliably. The methodology is based on an **observation-based theory of causation** that maps the observability of causal relationships in simulation traces to sets of basic causal assumptions. We show that certain combinations of causal assumptions have a negative impact on observability due to confounding effects. We further show that observability is best if only basic causal relationships involving *active events* and neither *omissions* nor *preventions* are taken into account[2]. From the observation-based theory of causation, we derive **two simple intervention rules** that form the core of the causal analysis methodology. From a conceptual point of view, the approach represents a combination of a counterfactual and an interventionist approach. The methodology is simple and can be easily integrated into an existing simulation framework.

The paper is structured as follows. In Section 2, we review some of the approaches for causal analysis. We argue that the conventional counterfactual approach to causal analysis by means of controlled experimentation is only suboptimal in the case of computational simulation since it does not do justice to the powerful features that distinguish an interactive computational simulation model from a physical phenomenon. Section 3 describes the basic idea of simulation-based causal analysis and motivates our intervention-based approach. An observational theory of causation, the resulting intervention rules, as well as their application to notoriously tricky cases of causation are described in Section 4. The practical integration of the causal analysis approach into a real simulation framework and its application is illustrated in Section 5.

2 CAUSAL ANALYSIS: A BRIEF OVERVIEW

The problem of causality has been discussed extensively throughout history. Hume's regularity theory marks a famous starting point in the formal treatment of causation [11]. According to it, "a cause is an object, followed by another, such that all objects similar to the first are followed by objects similar to the second. Or in other words, where the first object had not been, the second never had existed" [5]. Thus, two events A and B can be considered causally related (on the type level) if they always occur together and the assumed cause is always succeeded by the assumed effect (on the token level). Over the centuries, a number of criticisms against pure regularity theories for type causation have been raised. The following list summarises some of the well-known difficulties which gave rise to the development of alternative theories of causation [10].

[2] Due to the special treatment of omissions, our account is philosophically somewhat more closely related to Hitchcock's work on the deviant/default distinction [9] than to structural causal modelling [22] which considers omissions to be fully-fledged events in their own right.

Imperfect regularities: In many cases, causes invariably having to be followed by their effects is too strict a requirement. For example, if we only accepted this strict notion of regularity, then smoking would not be a cause of lung cancer. In other words, smoking causing lung cancer on the type level does not imply that every case of smoking has to be followed by a case of lung cancer on the token level.

Irrelevance: Not every condition A which is always followed by another condition B is necessarily relevant for B. For example, salt that has been hexed by a sorcerer invariably dissolves when placed in water. Nevertheless, hexing is clearly irrelevant for the dissolution of salt. In other words, hexing always being followed by dissolution on the token level does not imply that there is a token-causal relationship between hexing and dissolution.

Asymmetry: Causes always precede their effects but effects never precede their causes. Some theories based on pure regularity (e.g. strict correlation) may violate the asymmetry condition[3].

Spurious regularities: Event A invariably followed by B may indicate a causal relationship which, however, need not necessarily exist. For example, the crow of the rooster is regularly followed by the sunrise but it is clear that the rooster's crow does not *cause* the sun to rise. The source of spurious regularities are confounding events, i.e. common causes which are responsible for both A and B to happen together.

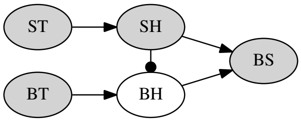
For a computational approach to causal analysis, the problem of asymmetry is easy to deal with: we just need to ensure that the stipulated cause happens temporally before the effect. As a response to the other problems, different alternative theories of causality have been developed. We review some of them in the following paragraphs.

Probabilistic causation: Instead of strictly requiring the common occurrence of causes and effects, the common assumption underlying probabilistic theories of causation is that *causes raise the probabilities of their effect.* Kleinberg and Mishra, whose work takes its inspiration from the work of Suppes [26] and Eells [6], give a probabilistic account of causality using temporal logics [13]. According to the authors, A is a prima facie cause of B if (i) A is reachable with a non-zero probability, and (ii) the probability of A being temporally followed by B is higher than the probability of just reaching B. This captures the idea of A raising the probability of B and satisfies the axioms of irregularity and asymmetry, yet it still leaves open the possibility for irrelevance and spurious regularities. To this end, Kleinberg and Mishra propose a hypothesis testing approach which helps to determine how significant each cause c is for its effect e by taking into account the *average difference in probabilities* $\epsilon_{avg}(c, e)$ that each c has on e. A problem with this approach is that the set of possible causes X needs to be known in order to calculate $\epsilon_{avg}(c, e)$. Furthermore, given the possibly vast size of X, testing may become a serious bottleneck. The heavyweight hypothesis testing step is necessary in this case since Kleinberg and Mishra aim to discover causal relationships in existing, *purely observational* data whose underlying causal structure is unknown and has to be inferred from the data using sophisticated statistical

[3] There are theories which state that causes are simultaneous and reciprocal with their causes [23]. This discussion is beyond the scope of this work.

Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle, had it not been preempted by Suzy's throw.

(a)



(b)

Fig. 2. Rock throwing scenario. Definition (a) and neuron diagram (b)

techniques. Although inferences about probabilities and conditional relationships between probabilities may certainly give insight into the causal structure of the observed process, it will always suffer from the problem that it cannot distinguish between causation and mere correlation.

Counterfactual causation: A popular approach to the analysis of causal relationships is through the idea of *counterfactual dependence*. Informally, a counterfactual statement is a conditional statement whose antecedent is hypothetical, i.e. contrary-to-fact. Consider, for example, the following statement: "if it had not rained yesterday, the street would not have been wet". The statement is hypothetical since it refers to a situation that has not occurred in reality (yesterday's not raining). Counterfactual theories of causation date back to the work of the American philosopher David Lewis [16]. He explained the semantics of counterfactuals by means of sets of *possible worlds*, i.e. states of affairs that *could have* been the case but *are not* the case in the actual world. According to this idea, statement "if A had happened then B would have happened" is true in the actual world if and only if B happens in some A-world (i.e. a possible world in which A happens) that is closer (i.e. more similar) to the actual world than any other A-world in which B does not happen. Causation can then be defined in terms of counterfactual dependence as follows: event A causes event B if and only if, had A not happened, B would not have happened either. We see that counterfactual causation is able to cope with the problem of irrelevance: had the salt not been hexed by a sorcerer, it would still have dissolved; therefore, hexing cannot be the cause of the dissolution. In the context of simulation, counterfactual approaches to causality seem particularly appealing. After all, each run of a probabilistic simulation model can be viewed as the realisation of a particular possible world[4]. By varying independent parameters of the model and re-running the simulation, counterfactual 'what-if' claims can be studied effectively. So it seems that Lewis' account of causality can be used right away to answer causal claims within a simulation study.

However, the critical aspect of Lewis' counterfactual theory is the notion of *similarity* between worlds mentioned above. When can a possible world W be considered 'more similar' or 'closer' to the actual world than another world W'? If a counterfactual theory is to be turned into an algorithm, those questions need to be answered. In order to understand the importance of similarity, consider the example in Figure 2 [21]. Intuitively, Suzy's throw was the cause of the bottle shattering. However, a naïve counterfactual analysis would not work: had Suzy not thrown the rock, the bottle would still have shattered because of Billy's throw. For illustration, it is helpful to

[4] Or sequence of possible worlds, depending on whether worlds are seen as *instant worlds* or as time-extended *worm worlds* [16].

introduce the idea of a *neuron diagram*. It was originally popularised by Lewis as a means of illustrating the temporal evolution of a particular set of causal relationships [16]. A neuron diagram is a directed graph in which nodes represent events ('neurons') and arrows represent causal relationships ('stimulatory connections') between events. Edges with dots in the end represent inhibitory connections or *preventions*. If a node is grey, then it is active and 'fires'; if it is white, then it is inactive. The temporal order is from left to right. A neuron diagram for the rock throwing example is shown in Figure 2b. Events are abbreviated as follows: *ST*= Suzy throws a rock, *BT* = Billy throws a rock, *SH* = Suzy's rock hits the bottle, *BH* = Billy's rock hits the bottle, and *BS* = the bottle shatters. We see that both Suzy's and Billy's rock hitting the bottle are causes of the bottle shattering (denoted by the arrows), respectively. However, the fact that Suzy's rock hits the bottle first prevents Billy's rock from hitting it (denoted by the edge with dot in the end).

The reason for the unsatisfying result of the counterfactual analysis is that we may have taken into account the wrong possible world(s): we looked at a world W in which A (e.g. Suzy's throw) has not happened, yet this not-happening triggered another background event A' (e.g. Billy's throw) which itself caused B. The original causal chain between A and B was thus *preempted* by the chain between A' and B which rendered the former undetectable. It is easy to see how similar cases of preemption could lead astray a naïve approach to counterfactual analysis in the virus propagation example given in the introduction. What is the problem? From an analysis point of view, world W that we looked at was *not similar enough* to the actual world since suddenly things happened — Billy's stone hitting the bottle — that did not happen in the actual world. The question now becomes how a sufficiently similar world can be found effectively. One possibility is to *generate* it through *active intervention*.

Intervention-based causation: As described above, in order to determine whether A causes B, it is not sufficient to merely observe the process and check whether A increases the probability of B. Instead, it has been proposed to check whether *doing* A, that is, *forcing A to happen*, increases the probability of B. These ideas can be attributed to Judea Pearl who argued that causality analysis is an inherently manipulatory task [22]. He likened conditional probabilities such as $Pr(B \mid A)$, i.e. the probability of B given that A, with $Pr(B \mid \mathbf{see}(A))$, i.e. the probability of B given that A has been *observed*. In order to check whether A has an influence on B, however, we need $Pr(B \mid \mathbf{do}(A))$, i.e. the probability of B given that A has been *actively forced to happen*. Changing the course of action by manipulating values intentionally (e.g. explicitly enforcing or inhibiting A), is called an *intervention*. Interventions are at the heart of controlled randomised experiments. By varying conditions between experimental and control groups, causal relationships can be identified. One problem is that, in purely observational studies, interventions are not possible. Since interventions cannot be performed retrospectively, causal relationships need to be derived from existing data sets in this case. This relates back to the notion of *counterfactual* statements described above. Whereas practical experiments involve *active intervention* (*'what happens if'*), counterfactual statements involve assumptions about *hypothetical intervention* (*'what would have happened if'*). Shpitser and Pearl argue that counterfactual statements are hard or even impossible to prove or disprove in purely physical experiments; after all, "we simply cannot perform an experiment where the same person is both given and not given treatment" [24]. Pearl found a solution to these problems. In his influential book, he describes a technique by means of which the effects of interventions can be studied by combining observational data and a *structural causal model*, a formal model of qualitative assumptions about causes and effects [22]. He was awarded the Turing Prize for his work on reasoning under uncertainty. Despite its power, defining a structural causal model for a given data set is a highly nontrivial task.

Interventions have also been proposed as a solution to the preemption issues in token causation introduced above. Consider again the rock throwing example. We saw that a simple counterfactual notion of causality is not successful since, had Suzy's throw not occurred, the bottle would still have shattered. Proposed solutions to this problem include Steve Yablo's idea of *de-facto dependence* [30] and Maudlin's idea of *sequential updating* [18]. De-facto dependence can be summarised as follows. E de-facto depends on C just in case had C not occurred, and had other suitably chosen factors been held fixed, then E would not have occurred [21]. 'Holding fixed' certain factors clearly represents an intervention. The critical bit is the second part of the sentence: what are the 'suitably chosen factors' that are to be held fixed? As nicely illustrated by Paul and Hall throughout their book, the problem has been debated extensively in the research community [21]. Sequential updating, on the other hand, can be roughly summarised as follows. In order to identify C as a cause of E given some other event B (henceforth referred to as *background factor*), we start by modifying the initial state of the world such that C does not occur. We then evolve forward until B is about to fire. Now, we tweak the state of the world such that B is prevented from firing. Further evolving forward, we observe that E does not happen, in which case C has been identified as the cause of E. 'Tweaking the state of the world' also describes an intervention. But again, it is unclear how exactly this intervention is to be made. The aim of this paper is to provide an answer to this question.

Let us summarise the ideas so far. Probabilistic accounts are practically useful, especially in the presence of (large quantities) of data, yet they suffer from the notorious causation-versus-correlation problem. Whilst it can be alleviated through sophisticated statistical analysis, it can never be fully circumvented. Furthermore, probabilistic accounts only work on the type level. Counterfactual theories of causation are conceptually strong, yet they leave open the question how the worlds that are 'similar enough' to the actual one and required for analysis should be found. Interventionist approaches propose to 'generate' those worlds through active manipulation, either by means of controlled experimentation or by extending the data with a structural causal model in a purely observational context. One problem with the latter approach is that the answers given by Pearl's approach are critically dependent upon the nature of the given structural causal model and the definition of an appropriate model is a highly difficult task. In the case of controlled experiments, it is unclear how the interventions should be designed such that cases of preemption are circumvented successfully.

Fortunately, computer simulation experiments are unlike both purely observational studies and real physical experiments. With a simulation model, we can generate as much data as we want. At the same time, we already have the 'cogs and wheels', the generator of the causal relationships that we are trying to uncover in the data, in our hand. If used appropriately, a simulation model allows us to travel back in time, change certain conditions and see *what would have happened if*. In the next sections, we describe an intervention-based methodology that detects cases of token causation effectively. We start with a formal view on simulation-based causal analysis in the next section.

3 SIMULATION-BASED CAUSAL ANALYSIS: BASIC IDEA

The methodology described in this paper is a computational version of the counterfactual approach to causal analysis described in Section 2, extended with Yablo's idea of de-facto dependence and Maudlin's idea of sequential updating. The goal of this section is to provide a high-level overview of the methodology. We start with a formal definition of simulation-based causal dependency in Section 3.1, followed by a description of the analysis methodology in Section 3.2.

3.1 A formal view on simulation-based causal dependency

We employ the Z notation for the formalisation [25]. Let $[Event]$ denote the set of all possible events that may occur within the context of the model (for example 'healthy' and 'infected' in the virus propagation model). A state $State == \mathbb{F}\, Event$[5] is then defined as a finite set of events (e.g. one for each agent in the simulation). A transition relation $Rel \subseteq State \times State$ is a relation between states. It describes the temporal evolution of the model by connecting each state with its possible predecessor and successor states. A trace $Trace == \mathrm{seq}\, State$ is then defined as a (possibly infinite) sequence of states. Consider again the rock throwing example. At each point in time, each of the five possible events (Suzy/Billy throwing, Suzy/Billy hitting, bottle shattering) can be either active or inactive and can thus be represented as a bit (active=1 and inactive=0). Since multiple events may happen at any point in time, each state is encoded as a list of bits, and a trace is encoded as a sequence of lists of bits. Figure 3a shows a possible trace. At time $t = 1$, both Suzy and Billy throw a rock ($ST = 1$ and $BT = 1$); at $t = 2$, Suzy's rock hits the bottle ($SH = 1$); at $t = 2$, the bottle shatters ($BS = 1$).

For clarity, we distinguish between a *simulation model* (a static representation of the scenario under consideration) and a *simulation procedure* used to *execute* the model by starting in a given initial state and traversing the state space according to the (mostly probabilistic) transition rules embedded in the model, resulting in one or more simulation traces. Formally, a simulation model $Model == (\mathbb{F}\, State, Rel)$ is defined as a tuple comprising a finite set of states (the state space) and a transition relation. Given a finite trace π, we denote with $\pi[i]$ the state at position i and with $\pi[-1]$ the final state. We further denote with $\#\pi$ the length of a trace. We can now define a *simulation procedure* Π as a function that returns for a given model m and a given initial state s_i all the traces starting in s_i and defined by the transition relation.

$$\Pi : Model \times State \to \mathbb{F}\, Trace$$

$$\forall\, m, s_i \bullet \Pi(m, s_i) = \{\langle s_0, .., s_n\rangle \mid s_0 = s_i \land \forall\, k : \{1 \,..\, n\} \bullet (s_{k-1}, s_k) \in Rel_m\}$$

Based on that, we can now give a trace-based definition of causal dependency.

[5] In the Z notation, == denotes definition and $\mathbb{F}\, Event$ denotes a finite set of elements of type *Event*.

$$C \rightarrow_{\pi} E \Leftrightarrow \; C \in \pi[0] \wedge \exists\, i : [1, \#\pi) \mid E \in \pi[i] \wedge \tag{1}$$

$$\exists\, \pi' \in \Pi(m, \pi[0] \setminus \{C\}) \mid \forall\, i : [0 \,..\, \#\pi') \bullet \pi'[i] \sim \pi[i] \wedge E \notin \pi'[i] \tag{2}$$

Given initial state s_i, trace $\pi : \Pi(m, s_i)$ of model m and events C and E happening in state 0 and $i > 0$, respectively, C causes E in π (denoted $C \rightarrow_{\pi} E$) if and only if (i) C happens in state 0 of trace π (henceforth referred to as the *actual trace*) and E happens in state $i > 0$, and (ii) had C not happened in state 0, then E would not have happened either in state *i ceteris paribus*, i.e. assuming everything else being equal (denoted $\pi[i'] \sim \pi[i]$). '$\sim$' thus represents a notion of *similarity* between traces that is central to the idea of de-facto dependence. Briefly anticipating the discussions further below, two traces are sufficiently similar for the purpose of causal analysis if and only if they differ with respect to the occurrence of events C at time t and E at times t' but not with respect to any other event. A detailed justification of this idea is given in Section 4. The formalisation also illustrates that simulation-based causal dependency is essentially an *existential quantification over possible traces*: C causes E in an actual trace π if and only if there exists a *counterfactual* trace π' that is similar enough to π which requires (among other factors, as described further below) that neither C nor E happens in the same state in which they happened in π. The formalisation suggests performing the counterfactual analysis by searching all traces produced by the simulation for counterfactual trace π'. In order for this to be possible, the set of traces produced by the simulation has to be small enough to be explored exhaustively, an assumption which is clearly unrealistic for most non-trivial simulation models. In the next section, we present a methodology that avoids exhaustive enumeration of all traces by *actively intervening* in the simulation process in order to *generate* a trace that is similar enough to the actual trace.

3.2 An intervention-based methodology for causal analysis

ALGORITHM 1: A typical simulation procedure

```
Require: initial state $s$, number of replications $r_{max}$, maximum time step $t_{max}$
$r \leftarrow 1$
$traces \leftarrow \langle\rangle$ {create empty list of traces}
while $r \leq r_{max}$ do
    $prng.seed(\cdot)$ {initialise PRNG}
    $t \leftarrow 1$
    $\pi \leftarrow \langle\rangle$ {create empty trace}
    while $t \leq t_{max}$ do
        $s' \leftarrow step(s, prng)$ {update simulation and produce new state}
        $\pi \leftarrow \pi \frown s'$ {append new state to trace}
        $t \leftarrow t + 1$
    end while
    $traces \frown \pi$ {append new trace to list of traces}
    $r \leftarrow r + 1$
end while
return traces
```

In practice, simulation procedure Π presented in the previous section will be typically realised as a procedural program that contains a set of nested loops. An example is shown in pseudo-code notation in Algorithm 1. Starting with an initial state, the simulation is executed a number of times using a pseudo-random number generator (PRNG). Each simulation run produces a trace of length t where t is the number of simulated time steps or ticks. The core of the procedure is an update function (called '*step*' in the example) which contains the actual (probabilistic) simulation logic. Let us now assume that a modeller uses this procedure to produce a set of traces. In one of those traces, denoted π, she observes event C in the initial state (i.e. $C \in \pi[0]$) and E in the final state (i.e. $E \in \pi[-1]$). For example, in the trace shown in Figure 3a, she observes Suzy's initial throw (thus $C = ST$) and the bottle's eventual shattering ($E = BS$). The token-causal analysis problem can now be stated as follows: **given trace π and events C and E, was C the cause of E in π?**

ALGORITHM 2: Simulation-based causal analysis procedure

```
Require: actual trace $\pi$, stipulated cause $C$, stipulated effect $E$, random seed used to produce
    actual trace $\pi$
Ensure: $C \in \pi[0], E \in \pi[-1]$
prng.init(seed) {use same PRNG seed in counterfactual as in actual run}
$s \leftarrow \pi[0] \setminus \{C\}$ {turn stipulated cause off in first state}
$t_{max} = \#\pi - 1$ {$t_{max}$ is equal to index of last element in actual trace}
$t \leftarrow 1$
while $t \leq t_{max}$ do
    $s' \leftarrow step(s, prng)$ {update simulation}
    $s \leftarrow intervene(s', \pi[t])$ {perform intervention}
    $t \leftarrow t + 1$
end while
return $\neg (E \in s)$ {$C$ causes $E$ iff $E$ does not happen in counterfactual trace}
```

As described in the previous section, answering this question requires finding a trace π' that is similar enough to the actual trace π and in which neither C nor E happens. Given the potentially vast number of traces, an exhaustive search is infeasible. Instead, we aim to generate the desired trace by intervening in the simulation process and, in doing so, 'steering' the simulation into the right direction from the perspective of analysis. This leads to the causal analysis procedure shown in Algorithm 2. Given a trace π in which both the stipulated cause C and the stipulated effect E happen, the procedure performs a counterfactual experiment based on the hypothesis that there is a causal relationship between C and E. In order to avoid effects of randomness in the simulation model, the first step of the counterfactual analysis procedure is to initialise the pseudo-random number generator (PRNG) with the same seed that was used to produce the actual trace π. In doing so, the counterfactual trace is kept as close to the actual one as possible. Before starting the simulation, C is turned off. The simulation is then run for a number of ticks (using the same update function as in the original simulation) and interventions are performed after each state update in order to modify the state accordingly[6]. When the simulation has finished and E is still present in the

[6] This represents essentially Maudlin's idea of *sequential updating* described in Section 2

final state, then the hypothesis that there is a causal relationship between C and E can be rejected; otherwise it is confirmed. The core of the causal analysis procedure is the intervention function (Line 7). It ensures that cases of preemption (such as those in the rock throwing example) are detected and not falsely reported as causal relationships. Consider the trace in Figure 3b. By simply turning off Suzy's throw (denoted by '1 → 0'), we allow a background event (Billy's rock hitting the bottle) to creep in. In the spirit of possible worlds semantics, we have thus moved into a world that differs from the original one in which both Suzy and Billy throw in three aspects (bolded in Figure 3b): (i) Suzy does not throw, (ii) Suzy does not hit, and (iii) Billy hits. The first difference is not critical since it represents the intervention itself (preventing Suzy from throwing). The second difference is a direct causal consequence of our intervention: if Suzy does not throw, then there is no chance for her rock hitting the bottle; it is thus also acceptable. The third difference is more critical since it represents a background event that was *enabled* by the intervention but not *caused* by it. It should thus not have occurred. For that reason, we need to perform a manual intervention in order to prevent Billy's rock from hitting the bottle. Interventions are the focus of the next section.

4 OBSERVING CAUSAL RELATIONSHIPS: FORMAL ANALYSIS

We assume that the simulation model under consideration is, in general, too complex to be analysed formally. For that reason, a controlled experiment needs to be conducted in order to restrict the traces produced by a simulation to those needed for counterfactual causal analysis. The controlled experiment consists of *active interventions* that aim to 'nudge' the simulation in such a way that only the trace that is strictly required for proving the counterfactual statement is generated. Starting from a modified initial state in which the stipulated cause is 'turned off', the trace is generated iteratively by observing and judging input–output relationships between events on the actual trace and manipulating it accordingly. The central questions that need to be answered here are: what types of interventions need to be made and what types of causal relationships can be detected? Are there different types of relationships (causal or non-causal) which, from an observer's perspective, are indistinguishable? The goal of this section is to derive answers to these questions in a formal and unambiguous way.

As illustrated in Section 2, a naïve approach to counterfactual causal analysis can be misleading and produce either *false negatives*, i.e. overlook relationships that are considered causal (as in the stone throwing example), or *false positives*, i.e. report relationships that are not to be considered causal in nature (examples are given further below). The problem is that the decision about whether a given relationship is to be considered causal is often dependent upon the context of the analysis, e.g. the problem domain under consideration. For example, in some cases it may make sense to consider an *omission*, i.e. a non-occurrence of an event, as a causal consequence of something else, whereas in other cases it may not. In some cases, it may make sense to consider the *prevention* of an event as a causal consequence of another event, whereas in other cases it may not. These conceptual difficulties are reflected in the vast amount of philosophical work on the metaphysics of causation, some of which has been briefly introduced in Section 2. As illustrated in the next paragraphs, trying to come up with a methodology that is capable of detecting all those different types of relationships is non-trivial. Rather than aiming for a unified account of causality itself, our ultimate goal is thus to provide a *parametrisable operationalisation* of the idea of counterfactual causation. By 'operationalisation', we mean an algorithmic procedure that takes a simulation model and performs a counterfactual analysis on it; by 'parametrisable' we mean that the algorithmic procedure can be configured such that it is capable of dealing with the different interpretations of what may count as a cause and what not, depending on the context of analysis.

The goal of this section is to clarify these informal ideas. We start with a formalisation of causation and prevention in Section 4.1. The question of whether omissions are allowed to be considered cause or effect in their own right is central to the type of causal relationship detectable by the analysis procedure and discussed in further detail in Section 4.2. As shown formally in Section 4.3, mixing different notions of causality has critical implications on the observability of causal relationships. If not dealt with appropriately, different causal and non-causal relationships become observationally indistinguishable and thus undetectable. The goal of the next section is to clarify the implications of certain manipulations on the occurrence or non-occurrence of other events and the resulting causal relationships.

4.1 Causation and prevention

We assume that the observable events produced by the simulation under consideration are recorded in a finite and temporally ordered trace $\pi = \langle \pi_0, \pi_1, \ldots \pi_{k-1} \rangle$ of length k. Each state π_i of a trace represents a finite set of events. We view those events as atomic entities, i.e. we are not interested in their internal structure.

*Definition 4.1 (**Occurrence and omission of events**).* We denote the occurrence of basic events with capital letters A, B, etc. We further denote the omission, i.e. non-occurrence, of basic events with $\neg A$,$\neg B$, etc.

Since we are focussing on token causation and we consider the trace under observation as a deterministic outcome of a non-deterministic simulation model, we employ, for a single trace, a notion of causality that is based on necessity and sufficiency. Consider again the stone throwing example given above. If we could replay history and prevent Suzy's toss from happening *ceteris paribus* (other things being equal!), we would expect the bottle not to shatter — despite Billy's toss. This is the case because the *ceteris paribus* rule would prevent Billy's stone from hitting and thus shattering the bottle[7]. As a consequence of this assumption, if two events A and B are causally related in a purely deterministic trace, they are both necessary and sufficient for each other. In other words, in a perfectly deterministic world without side effects, if A causes B, then, if A does not happen, B will not happen either; if B does not happen, A cannot have happened before. A causing B in a deterministic world thus implies A being both necessary and sufficient for B. This leads to the following definition.

[7] Note that this assumption only holds if there is a true causal relationship between the toss of the stone and the shattering of the bottle. Furthermore, as described in Section 3, the *ceteris paribus* principle is central to this idea.

*Definition 4.2 (**BC: Basic Causation**).* Let π be a simulation trace. In order for event A to cause event B (denoted $A \rightarrow B$, by analogy with the notation of neuron diagrams), A's occurrence has to be both sufficient and necessary for B in π[8].

$$A \rightarrow B \equiv (A \Rightarrow B) \wedge (B \Rightarrow A) \equiv A \Leftrightarrow B \tag{3}$$

It is important to note that, although $A \Leftrightarrow B \equiv \neg\, A \Leftrightarrow \neg\, B$, this does not imply that $\neg\, A \Leftrightarrow \neg\, B \equiv \neg\, A \rightarrow \neg\, B$. In order for the latter to hold, we need to explicitly accept omissions as both causes and effects, as described in Sections 4.2.1 and 4.2.2. It is further important to note that Definition 4.2 does not violate the principle of asymmetry mentioned in Section 2, although the logical equivalence may suggest that. Since we assume that all events described by capital letters are recorded in a simulation trace, there is a natural temporal ordering between them. This ordering ensures temporal asymmetry between events, notwithstanding their logical equivalence.

From Definition 4.2, it follows that causation is transitive, as shown below.

Theorem 4.3 (**TOC: Transitivity Of Causation**). *If A causes B and B causes C then A causes C.*

Proof.

$$\dfrac{\dfrac{A \rightarrow B}{A \Leftrightarrow B}\,[\text{BC}] \qquad \dfrac{B \rightarrow C}{B \Leftrightarrow C}\,[\text{BC}]}{\dfrac{A \Leftrightarrow C}{A \rightarrow C}\,[\text{BC}]}\,[\text{Def. of} \Leftrightarrow]$$

□

In addition to the notion of an event A causing another event B, we further allow for an event A to *prevent* another event B from occurring, as described below.

*Definition 4.4 (**PREV: Prevention**).* In order for event A to prevent event B (denoted $A \multimap B$, by analogy with the notation of neuron diagrams), A's occurrence has to be both necessary and sufficient for B's omission.

$$A \multimap B \equiv (A \Rightarrow \neg\, B) \wedge (\neg\, B \Rightarrow A) \equiv A \Leftrightarrow \neg\, B \tag{4}$$

By analogy with Definition 4.2 (basic causation), although $A \Leftrightarrow \neg\, B$ is equivalent to $\neg\, A \Leftrightarrow B$, this does not imply that $\neg\, A \Leftrightarrow B$ is equivalent to $\neg\, A \multimap \neg\, B$. In order for the latter to hold, we need to explicitly accept omissions as both causes and effects, i.e. to allow omissions to appear either before or after the '$\rightarrow$' operator. The question whether omission should count as causes or effects is the subject of the next section.

4.2 Dealing with omissions

An omission is defined as an event that is not happening. This provokes the question whether omissions may themselves act as causes or effects (or both) and whether they should be treated uniformly with active events. Consider, for example, a situation in which Alice is about to drown in a lake. If Bob does not help, is Bob's not helping the cause of Alice's death? It may certainly count as a *contributing* cause. In the same spirit,

[8] A is sufficient for B if and only if B always occurs whenever A occurs, denoted $A \Rightarrow B$. A is necessary for B if and only if A always occurs whenever B occurs, denoted $B \Rightarrow A$. Sufficiency and necessity together yield *equivalence*, denoted $A \Leftrightarrow B$.

if Bob drags Alice along to a bar the evening before an important exam and Alice fails the exam, did Bob cause Alice not to pass? Is 'not passing' a valid causal consequence of the trip to the bar? Answers to these questions have important consequences on the formal nature of causality, as described further below.

The treatment of omissions has been debated extensively in the literature [9, 20, 29]. Within the framework of structural causal models, omissions are viewed as ordinary events. For example, Billy's not throwing the stone is simply represented as $BT = 0$, i.e. as the assignment of one of a range of permitted values to a causal variable. From the perspective of the formalism, there is thus no conceptual difference to an active event such as $BT = 1$. Several authors have pointed out that such a uniform treatment of active events and omissions may be problematic. For example, Paul and Hall introduced two scenarios (one involving overdetermination and one involving bogus prevention) whose structural causal models are isomorphic despite there being an obvious difference in their causal interpretation [21]. As a consequence, they argue, there must be a non-structural difference between those scenarios. Hitchcock provides a possible explanation by introducing the notion of *default* and *deviant* events. The rationale behind this distinction is nicely summarised in the *Principle of Sufficient Reason (POSR)* which states that "when a set of variables all take their default value, they cannot by themselves cause another variable to take a deviant value" [9]. Or, as Kahneman and Miller put it, "a cause must be an event that could easily have been otherwise. In particular, a cause cannot be a default value among the elements that the event X has invoked." [12]. Using the default/deviant distinction, Hitchcock introduces the notion of a *self-contained network* in which "it is never necessary to leave or augment the network in order to explain why the variables within the network take the values that they do." [9]. In other words, a causal network between C and E is self-contained if for all variables V (excluding C) holds that, if all other variables $V' \neq V$ (including C) take their default values and all off-path values keep their actual values, V also takes its default value. Hitchcock concludes that if the network between C and E is self-contained and E is counterfactually dependent upon C, then C causes E. We show in the following sections that a uniform treatment of omissions and active events is problematic from an observational perspective as it masks certain true causal relationships. This indicates that omissions may, in fact, have to be treated differently from active events.

4.2.1 Omissions as causes. Can an omission be a cause of another event? In other words, can 'not doing A' be a cause of B? If we accept this idea, then we have to accept the following axiom which we refer to as *causation by omission.*

*Definition 4.5 (**CBO: Causation By Omission**).* If A's omission is both necessary and sufficient for B's occurrence and we accept that omissions are allowed to be causes then A's omission can be considered to cause B's occurrence.

$$\frac{\neg A \Leftrightarrow B}{\neg A \rightarrow B}$$

Adding this axiom to the previous ones does not by itself make any difference since we cannot prove any further theorems. However, in Section 4.3 we show that adding CBO complicates purely observation-based causal analysis significantly since it may produce false positives. In the next section, we investigate whether an omission may count as an effect, i.e. a causal consequence of another event.

4.2.2 Omissions as effects. Can an omission be an effect of another event? In other words, can 'B's not happening' be a causal consequence of A? If we accept this idea, then we have to accept the following axiom which we refer to as *causation by prevention.*

*Definition 4.6 (**CBP: Causation By Prevention**).* If A's occurrence is both necessary and sufficient for B's omission and we accept that omissions are allowed to be effects then A's occurrence can be considered to cause B's omission. From Definition 4.4, it follows that causing a non-occurrence is equivalent to prevention and we thus refer to this axiom as causation by prevention.

$$\frac{A \Leftrightarrow \neg B}{A \rightarrow \neg B}$$

If we accept omissions to be effects, i.e. causal consequences, then we also have to accept the fact that, if A causes B which, in turn, prevents C then A can be considered to prevent C, as proven below.

THEOREM 4.7 (**PCP: PREVENTION BY CAUSED PREVENTION**). *If we accept TOC and CBP then it follows that, if A causes B and B prevents C, then A prevents C.*

PROOF.

$$\dfrac{\dfrac{A \rightarrow B \qquad \dfrac{B \multimap C}{B \rightarrow \neg C}\,[\text{CBP}]}{A \rightarrow \neg C}\,[\text{TOC}]}{A \multimap C}\,[\text{PREV}]$$

□

No further theorems can be proven by adding CBP as an axiom. However, by analogy with CBO described above, accepting CBP complicates counterfactual analysis significantly, as described in further detail in Section 4.3. With the formal definitions of causation, prevention, causation by omission, and causation by prevention, we can now investigate what types of causal and non-causal relationships can be detected using counterfactual analysis.

4.3 An observation-based theory of causation

The basic idea of simulation-based counterfactual analysis is that of a controlled experiment. This means that the simulation is run twice: once with the potential cause C (the 'control variable') activated and once with C deactivated. C can then be considered a cause of E (the 'dependent variable') if and only if E only happens in the first run, but not in the second. The rock throwing example nicely illustrates that a naïve approach is doomed to fail: preventing Suzy from throwing the rock in the beginning does not solve the problem since it allows a background event (Billy's rock hitting the bottle) to creep in. In the spirit of possible worlds, the initial intervention produced a world that is too far away from the one required for counterfactual analysis. In general, what we need is a world in which the following requirements are satisfied:

(1) The potential cause C is turned off.
(2) Due to (1), none of the causal consequences of C happen.

Table 1. The four different observations for a trace of length 2 with causally unrelated events

		Observation		
Case	Causal assumption	original trace	manipulated trace	Req. axioms
0.1	$C \not\twoheadrightarrow E$	$\langle\{C\}, \{E\}\rangle$	$\langle\{\neg\, C\}, \{E\}\rangle$	BC
0.2	$C \not\twoheadrightarrow \neg\, E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBP
0.3	$\neg\, C \not\twoheadrightarrow E$	$\langle\{\neg\, C\}, \{E\}\rangle$	$\langle\{C\}, \{E\}\rangle$	BC+CBO
0.4	$\neg\, C \not\twoheadrightarrow \neg\, E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP

(3) Background events and their causal consequences do not happen.

Simply turning off the stipulated cause (Suzy's throwing) does not produce such a world since background effects may creep in. Instead, following Pearl's suggestions, we need to perform manual interventions during the simulation. That means that, if necessary at some point t during the simulation, a manual manipulation of the state is performed before the simulation is allowed to continue. This directly corresponds with Yablo's idea of tweaking the state of the world before evolving forward in time.

In order to come up with the rules for performing such interventions, let us first consider all possible cases that might occur. Remember that, according to the definition of causation given in Section 4.1 above, causes are considered necessary and sufficient for their effects within a single trace. Let $\pi = \langle \pi_0, \pi_1 \rangle$ be a simple trace comprising only two states. We assume that event C is the only event that is observed to happen in the first state, i.e. $\pi_0 = \{C\}$ and E is the only event that is observed to happen in the second state, i.e. $\pi_1 = \{E\}$. We can now investigate the different causal or non-causal relationships that might exist between C and E and what implications they have on the four different observations that can be made when running a counterfactual experiment.

4.3.1 One cause, one effect.

We start with all the cases where there is no causal relationship between C or $\neg\, C$ and E or $\neg\, E$, respectively.

Case 0.1 ($C \not\twoheadrightarrow E$): In the first case, we let $\pi = \langle\{C\}, \{E\}\rangle$ and assume that there is no causal relationship between C and E (denoted by $C \not\twoheadrightarrow E$). As a consequence, if we turn off C, then we expect E not to change to $\neg\, E$. So we can conclude that an external observer will either see $\langle\{C\}, \{E\}\rangle$ (the original trace) or $\langle\{\neg\, C\}, \{E\}\rangle$ (the manipulated trace).

Case 0.2 ($C \not\twoheadrightarrow \neg\, E$): In the second case, we let $\pi = \langle\{C\}, \{\neg\, E\}\rangle$ and assume that there is no causal relationship between C and E's omission. As a consequence, we observe $\langle\{C\}, \{\neg\, E\}\rangle$ as the original trace; if we turn off C, then we observe $\langle\{\neg\, C\}, \{\neg\, E\}\rangle$. Note that, in order to be able to justifiably talk about C not causing $\neg\, E$, we need to accept that there *might have been* a causal relationship between an active event and an omission which is only the case if we accept CBP as an axiom.

Case 0.3 ($\neg\, C \not\twoheadrightarrow E$): In the third case, we let $\pi = \langle\{\neg\, C\}, \{E\}\rangle$ and assume that there is no causal relationship between $\neg\, C$ and E. As a consequence, we observe $\langle\{\neg\, C\}, \{E\}\rangle$ as the original trace and $\langle\{C\}, \{E\}\rangle$ as the manipulated trace. Similar to the previous case, in order to be able to justifiably talk about $\neg\, C$ not causing E, we

Table 2. The four different observations for a trace of length 2 with one cause and one effect

		Observation		
Case	Causal assumption	original trace	manipulated trace	Req. axioms
1.1	$C \rightarrow E$	$\langle\{C\}, \{E\}\rangle$	$\langle\{\neg C\}, \{\neg E\}\rangle$	BC
1.2	$C \rightarrow \neg E$	$\langle\{C\}, \{\neg E\}\rangle$	$\langle\{\neg C\}, \{E\}\rangle$	BC+CBP
1.3	$\neg C \rightarrow E$	$\langle\{\neg C\}, \{E\}\rangle$	$\langle\{C\}, \{\neg E\}\rangle$	BC+CBO
1.4	$\neg C \rightarrow \neg E$	$\langle\{\neg C\}, \{\neg E\}\rangle$	$\langle\{C\}, \{E\}\rangle$	BC+CBO+CBP

need to accept that there *might have been* a causal relationship between an omission and an active event which is only the case if we accept CBO as an axiom.

Case 0.4 ($\neg C \nrightarrow \neg E$): In the fourth and final case in this group, we let $\pi = \langle\{\neg C\}, \{\neg E\}\rangle$ and assume that there is no causal relationship between C's omission and E's omission. As a consequence, we observe $\langle\{\neg C\}, \{\neg E\}\rangle$ as the original trace and $\langle\{C\}, \{\neg E\}\rangle$ as the manipulated trace. By analogy with the previous two cases, we need to accept *both* CBO *and* CBP as axioms for this case to be conceivable.

A summary of the previous cases is given in Table 1. We now revisit the four cases described above but assume that there *is* a causal relationship between the two respective events.

Case 1.1 ($C \rightarrow E$): In the first case, we assume that C does, in fact, cause E. If we turn off C, then E will be deactivated as well. So we can conclude that we observe either $\langle\{C\}, \{E\}\rangle$ (the original trace) or $\langle\{\neg C\}, \{\neg E\}\rangle$ (the manipulated trace).

Case 1.2 ($C \rightarrow \neg E$): In the second case, we assume that C causes E's omission. In other words, C prevents E. As the original trace, we observe $\langle\{C\}, \{\neg E\}\rangle$; if we turn off C, then we observe $\langle\{\neg C\}, \{E\}\rangle$. Note that this case is valid only if we accept that prevention can be considered an actively causal activity, i.e. if we accept CBP.

Case 1.3 ($\neg C \rightarrow E$): In the third case, we assume that C's omission causes E. Here, we observe $\langle\{\neg C\}, \{E\}\rangle$; as the manipulated trace, we observe $\langle\{C\}, \{\neg E\}\rangle$. We can see that, from an observer's perspective, this case is symmetrical to Case 1.2. Note that this case is valid only if we accept that omissions can be true causes, i.e. if we accept CBO.

Case 1.4 ($\neg C \rightarrow \neg E$): In the fourth and final case, we assume that C's omission causes E's omission. As the original trace, we observe $\langle\{\neg C\}, \{\neg E\}\rangle$; as the manipulated trace, we observe $\langle\{C\}, \{E\}\rangle$. From an observer's perspective, this case is symmetrical to 1.1. Note that this case is valid only if we accept *both* CBO *and* CBP.

The results of the previous four cases are summarised in Table 2. At this point, the different observations are still distinguishable, i.e. there is a clear one-to-one mapping between the causal assumption and the observation. We will see in the next section that the situation gets worse when we consider the presence of background events.

4.3.2 Two causes, one effect. In the previous examples, we assumed that C and E were the only events in the example trace. We will now allow for one additional event D to happen in state π_0 and investigate what implications that has on the observations

Table 3. The 16 different observations for a trace of length 2 with two causes and one effect

Case	Causal assumption	Observation		Req. axioms
		orig. trace	manip. trace	
2.1	$C \rightarrow E$ and $D \rightarrow E$	$\langle\{C\}, \{E\}\rangle$	$\langle\{\neg\, C\}, \{E\}\rangle$	BC
2.2	$C \rightarrow E$ and $D \rightarrow \neg\, E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBP
2.3	$C \rightarrow \neg\, E$ and $D \rightarrow E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{E\}\rangle$	BC+CBP
2.4	$C \rightarrow \neg\, E$ and $D \rightarrow \neg\, E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBP
2.5	$\neg\, C \rightarrow E$ and $D \rightarrow E$	$\langle\{\neg\, C\}, \{E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO
2.6	$\neg\, C \rightarrow E$ and $D \rightarrow \neg\, E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP
2.7	$\neg\, C \rightarrow \neg\, E$ and $D \rightarrow E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{E\}\rangle$	BC+CBO+CBP
2.8	$\neg\, C \rightarrow \neg\, E$ and $D \rightarrow \neg\, E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP
2.9	$C \rightarrow E$ and $\neg\, D \rightarrow E$	$\langle\{C\}, \{E\}\rangle$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBO
2.10	$C \rightarrow E$ and $\neg\, D \rightarrow \neg\, E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP
2.11	$C \rightarrow \neg\, E$ and $\neg\, D \rightarrow E$	$\langle\{C\}, \{\neg\, E\}\rangle$	$\langle\{\neg\, C\}, \{E\}\rangle$	BC+CBO+CBP
2.12	$C \rightarrow \neg\, E$ and $\neg\, D \rightarrow \neg\, E$	$\langle\{C\}, \{\neg\ \}\rangle E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP
2.13	$\neg\, C \rightarrow E$ and $\neg\, D \rightarrow E$	$\langle\{\neg\, C\}, \{E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO
2.14	$\neg\, C \rightarrow E$ and $\neg\, D \rightarrow \neg\, E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP
2.15	$\neg\, C \rightarrow \neg\, E$ and $\neg\, D \rightarrow E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{E\}\rangle$	BC+CBO+CBP
2.16	$\neg\, C \rightarrow \neg\, E$ and $\neg\, D \rightarrow \neg\, E$	$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	$\langle\{C\}, \{\neg\, E\}\rangle$	BC+CBO+CBP

Table 4. Possible observations and different combinations of causal axioms entailing them

Observation	BC+CBO+CBP	BC+CBO	BC+CBP	BC
$\langle\{C\}, \{E\}\rangle$ and $\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	1.1, 2.9	1.1, 2.9	1.1	1.1
$\langle\{C\}, \{E\}\rangle$ and $\langle\{\neg\, C\}, \{E\}\rangle$	0.1, 2.1	0.1, 2.1	0.1, 2.1	0.1, 2.1
$\langle\{C\}, \{\neg\, E\}\rangle$ and $\langle\{\neg\, C\}, \{\neg\, E\}\rangle$	0.2, 2.2, ...		0.2, 2.2, 2.4	
$\langle\{C\}, \{\neg\, E\}\rangle$ and $\langle\{\neg\, C\}, \{E\}\rangle$	1.2, 2.3, 2.11		1.2, 2.3	
$\langle\{\neg\, C\}, \{E\}\rangle$ and $\langle\{C\}, \{\neg\, E\}\rangle$	1.3, 2.5, 2.13	1.3, 2.5, 2.13		
$\langle\{\neg\, C\}, \{E\}\rangle$ and $\langle\{C\}, \{E\}\rangle$	0.3	0.3		
$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$ and $\langle\{C\}, \{\neg\, E\}\rangle$	0.4, 2.6, 2.8, ...			
$\langle\{\neg\, C\}, \{\neg\, E\}\rangle$ and $\langle\{C\}, \{E\}\rangle$	1.4, 2.7, 2.15			

made when performing a counterfactual analysis. We still assume that the focus of analysis is on the causal relationship between C and E and that, as a consequence, the manipulation made is to turn off C. Also note that we only consider cases in which there is a causal relationship of some sort between *both* C and E *and* D and E. The reason is that cases in which only one of the two events has a causal relationship to E reduces to one of the four cases described in Section 4.3.1 above. Finally, in order to be consistent with the semantics of neuron diagrams, we consider preventions to be stronger than causal relationships. That is, if $C \rightarrow E$ and $D \multimap E$ at the same time, we assume that $\multimap$ 'wins' and E gets inhibited.

For brevity, we do not give full descriptions of all sixteen cases here, a summary is provided in Table 3. We can see that the ambiguity with respect to inferences that can be drawn from the observations is considerable. From an observer's perspective, cases 1.1 and 2.9 are indistinguishable, as are cases 0.1 and 2.1, cases 1.3, 2.5, and 2.13, etc. The overlap becomes clearly apparent when we map the observations to the cases that they may represent, as shown in Table 4 (for space limitations, some cases have been omitted). We can see that the ambiguity is greatest when the notions of BC, CBO

Event	t1	t2	t3
ST	1	0	0
BT	1	0	0
SH	0	1	0
BH	0	0	0
BS	0	0	1

(a)

t1	t2	t3
1→0	0	0
1	0	0
0	**0**	0
0	**1**	0
0	0	1

(b)

t1	t2	t3
1→ 0	0	0
1	0	0
0	**0!**	0
0	**1→0**	0
0	0	0

(c)

t1	t2	t3
1	0	0
1 → 0	0	0
0	1	0
0	0	0
0	0	1

(d)

Fig. 3. Simulation trace for the rock throwing scenario with both Suzy and Billy throwing (a), the counterfactual case where only Billy throws (b), the counterfactual case with intervention where Billy throws but does not hit the bottle (c), and the counterfactual case where Suzy throws and hits the bottle (d).

and CBP are mixed, followed by BC and CBO, and BC and CBP. The ambiguity is smallest when only BC is used. The only overlap is between cases 0.1 and 2.1 where the latter (a case of preemption) could be mistaken for a non-causal relationship or vice versa. The particular problem in case 2.1 is that the trace is too coarse-grained (i.e. consists of two events only), as a consequence of which the preemption effect remains undetectable. We show in the next section that if the trace is fine-grained enough then preemption effects become, in fact, detectable.

The analysis in the previous paragraphs indicates that viewing omissions as causes is problematic from an observational perspective as it masks true causal relationships. Bridging the gap to Hitchcock's work, we may thus consider the non-occurrence of an event as its default state and the occurrence as deviant. According to the Principle of Sufficient Reason, the latter may well act as a cause, whereas the former would not.

4.4 Intervention rules

The formalisation of causality and observability given in the previous sections provides the basis for the intervention rules. Given Table 4, when progressing a computational model step-by-step and observing the activation and deactivation of events, we consider an event E to be a causal consequent of another event C if and only if, had C not occurred (*ceteris paribus*), E would not have occurred either. Remember that the *ceteris paribus* rule consists of two parts: (i) controlling for randomness and (ii) preventing confounding effects. As described in Section 3.2, (i) can be achieved practically by seeding the pseudo-random number generator (PRNG) in the counterfactual experiment with the same value that was used to produce the actual trace. In doing so, undesired effects of randomness are eliminated and the two traces are kept as closely aligned as possible. For (ii), the analysis in the previous sections yield the following two intervention rules that form the core of the causal analysis approach.

Rule 1: If, as a consequence of turning off a potential cause C, some other event D *starts* happening in the counterfactual trace (w.r.t. the actual trace), then D must be deactivated before continuing the simulation (since D cannot have been caused by C but only by a background event, as described above)[9].

[9] It is important to note D may well be the same event as C, just happening at a different point in time. For example, in the actual trace, the bottle may shatter ($BS = 1$) at $t = 3$ whereas in the counterfactual trace it may shatter at $t = 5$. Despite the events themselves being equivalent, $BS = 1$ in the counterfactual

Rule 2: If, as a consequence of turning off a potential cause C, some other event D *stops* happening in the counterfactual trace (w.r.t. the actual trace), the simulation can be continued without intervention (since D must have been caused by C, as described above).

The application of the two rules to the rock throwing example results in the trace shown in Figure 3c. Here, event $BH = 1$ which is caused by $BT = 1$ at time $t = 2$ is manually turned off as instructed by Rule 1 (denoted by '$1 \rightarrow 0$'). As instructed by Rule 2, $SH = 0$ is left unchanged (denoted by '0!'). Despite Billy's still throwing at time $t = 0$, the causal background chain resulting from that event is now prevented from occurring. As a consequence of the rule application, BS no longer happens at time $t = 3$, correctly suggesting that C is in fact the cause of E. If we perform the same analysis for Billy's throw as shown in Figure 3d, the result is that the bottle still shatters which correctly rules out Billy's throw as a cause. The definition of causation against the background of simulation-based analysis can now be given as follows.

Definition 4.8. C can be considered a cause of E if and only if the following two conditions are satisfied.

(1) In a simulation run in which C happens, E also happens at a later point in time.
(2) In a simulation run in which C is turned off and intervention rules 1 and 2 described above are applied, E does not happen at a later point in time.

Function *intervene* used in Algorithm 2 can now be defined as follows.

$$\mathit{intervene} : \mathit{State} \times \mathit{State} \rightarrow \mathit{State}$$

$$\forall\, s', s : \mathit{State} \bullet \mathit{intervene}(s', s) = s' \setminus s$$

Here, the counterfactual new state s' is modified such that all events that started occurring in the counterfactual case but have not occurred in the actual case are suppressed (i.e. removed from the set of active events) according to Rule 1. Since Rule 2 does not require any active changes to a state, it is not represented explicitly. As a consequence, a modified counterfactual trace π' differs from the actual trace π in that every state s' in π' is a subset of the related one in π, i.e. $\forall\, s_i : \pi, s'_i : \pi' \bullet s'_i \subseteq s_i$. Based on that, we can now give a refined definition of the similarity relation $\sim\, \subseteq \mathit{Trace} \times \mathit{Trace}$ between traces according to which $\pi \sim \pi' \Leftrightarrow \forall\, i : [0 \,..\, \#\pi] : \pi'[i] \subseteq \pi[i]$. The similarity relation thus defines a *partial order* on the set of traces.

With the rock throwing scenario, an infamous problem of preemption has been shown above to be circumvented successfully using this approach. In the next section, we revisit some of the other frequently occurring examples in the literature and investigate how far the methodology is able to deal with them.

4.5 Evaluation: dealing with tricky situations

In this section, we analyse a number of examples from the literature that have been shown to pose challenges to different causality accounts [21]. Most of them involve some sort of preemption and contain multiple competing and interacting causal chains.

trace needs to be suppressed since it happens at a different point in time than in the actual trace and thus represents a difference that is to be eliminated. An example can be found in Section 5.

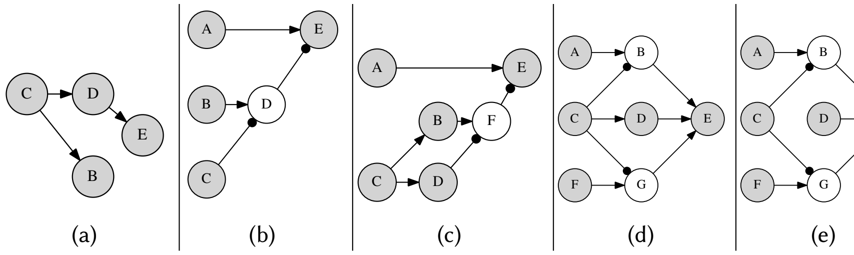


Fig. 4. Neuron diagrams for the five evaluation scenarios

	t1	t2	t3
C	**0**	0	0
D	0	0	0
B	0	0	0
E	0	0	0

(a)

	t1	t2	t3
A	1	0	0
B	1	0	0
C	**0**	0	0
D	0	**1→0**	0
E	0	0	1

(b)

	t1	t2	t3	t4
A	1	0	0	0
C	**0**	0	0	0
B	0	**0!**	0	0
D	0	**0!**	0	0
F	0	0	0	0
E	0	0	0	1

(c)

	t1	t2	t3
A	1	0	0
C	**0**	0	0
F	1	0	0
B	0	**1→0**	0
D	0	**0!**	0
G	0	**1→0**	0
E	0	0	0

(d)

	t1	t2	t3
A	1	0	0
C	**0**	0	0
F	1	0	0
B	0	**1→0**	0
D	0	1	0
G	0	**1→0**	0
E	0	0	1

(e)

Fig. 5. Simulation traces for the five evaluation scenarios

The examples are increasingly complex and have different requirements with respect to the application of Rules 1 and 2.

Example 1: Joint effects (no intervention required). Let us first consider a very simple scenario involving joint effects of a common cause shown in Figure 4a. Here, event *C* triggers both event *B* and event *D*. Event *D* itself triggers event *E*. *C* is thus a cause of *E* whereas *B* is not. However, the interesting aspect of this example is that, despite not being a cause, *B* is both necessary and sufficient for *D*. There is a spurious relationship between *B* and *D* that is not causal in nature and that a pure regularity account of causality is not able to detect. However, the causal relationship between *C* and *E* can easily be determined by means of a simple counterfactual analysis as shown in Figure 5a: if we prevent *C* from happening, then *E* does not occur. If we do the same analysis with *B*, then we observe that *E* still happens which correctly rules out *B* as a cause of *E*. So, in this simple example, neither Rule 1 nor 2 need to be applied.

Example 2: Double prevention (application of Rule 1). The next example involves a case of *dependence by double prevention* (see Figure 4b). Here, *A* is clearly the cause of *E*, *B* starts a process that aims to prevent *E* from happening, and *C*, in turn, starts a process that aims to prevent *B*'s prevention process. A simple analysis for counterfactual dependence between *A* and *E* would correctly identify *A* as a cause of *E* (if we turn off *A* then *E* becomes inactive, too). The same analysis would also correctly rule out *B* as a cause of *E* since *E* is not counterfactually dependent upon *B*. The problem is that a similar analysis for dependence between *C* and *E* would identify *C* as a cause of *E* which is clearly wrong! To see this, consider a situation in which *C* is turned off. In this case, *D* will be activated and prevent *E* from happening. *E* thus counterfactually depends on *C* without *C* being a cause of *E*. So, in order to prevent *B* from activating *D* after turning off C, we need to apply Rule 1, as shown in Figure 5b. Now, C is correctly ruled out as a cause of E since E still happens.

Example 3: Early preemption with threat and counteract (application of Rule 2). The next example (shown in Figure 4c) represents a more involved case of early preemption. Here, A is clearly the cause of E. However, the causal process between A and E is threatened to be prevented by the process running from C through E via B. In order to make things even more complicated, the threat itself is counteracted by the process from C through E via D. A would be easily identified as a cause of E by means of counterfactual analysis. The same analysis would also correctly rule out C as a cause of E, however merely by accident! The reason is that the result of the counterfactual analysis critically depends on which facts are held fixed. As described by Paul and Hall [21], some de-facto dependency accounts falsely report E to be counterfactually dependent upon C. For example, if we, according to Hitchcock's account [8], turn off C while only holding the value of B fixed (i.e. active), E will become inactive. The analysis would thus falsely report E to be de-facto dependent on C and thus a causal consequence of it. In order to avoid this, Rule 2 has to be applied to *both B and D*. That is, if we turn off C, we need to make sure that neither B nor D fires in order to prevent F from firing and inhibiting E. As shown in Figure 5c, Rule 2 guarantees that C is ruled out as a cause of E.

Example 4: Symmetrical overdetermination (application of Rule 1 and 2). Preemption of the type present in the previous examples is often referred to as *asymmetric overdetermination*. The reason is that, despite there being two (or more) sufficient and necessary causes for an effect, only one of them actually happens. The next example involves a case of *symmetrical overdetermination*. Here, multiple events that are both necessary and sufficient for an effect E to happen exist. The neuron diagram of the example is shown in Figure 4d. Here, C is clearly the cause of E. However, a naïve counterfactual analysis would fail to detect that. For, had C not fired, B and G would have fired, both of which would have activated E. In that case, B and G would become *symmetrically redundant causes* of E. In order to correctly identify C as a cause, we thus need to prevent B and G from firing after turning off C which is done by applying Rule 1 at time $t = 2$. However, we also need to apply Rule 2 in order to keep D silent after C has been turned off and thus prevent it from activating E itself. As shown in Figure 5d, if both rules are applied, C is correctly identified as a cause of E. The importance of Rule 2 for this example can be highlighted by slightly modifying the scenario such that the causal connection between C and E disappears, as shown in Figures 4e and 5e. If we compare the two scenarios, we see that E is not counterfactually dependent upon C in both cases. A counterfactual analysis would thus rule out C as a cause of E in both cases — in the first case falsely, in the second case correctly. Even when applying Rule 1 in both cases, the difference remains undetected. It is precisely the application of Rule 2 in the first case that makes the crucial difference.

5 EXAMPLE: A VIRUS PROPAGATION MODEL IN PYTHON MESA

In order to illustrate the practical application of the causal analysis methodology described in the previous section, we now return to the virus propagation example introduced in Section 1 and implement it in Python Mesa, a publicly available agent-based modelling framework[10]. We further integrate into the model the capability to perform causal analyses based on the methodology described in Section 3. The integration is very simple and only requires minor changes to the model. An integration into other frameworks such as Repast or Mason can, of course, be achieved accordingly.

As described in Section 4.4, an intervention is essentially a mapping from the state in the actual trace and the state in the current trace to a new manipulated state. As a consequence, we implement the intervention in Mesa as a function object (also known as *functor*) that stores a reference to the actual trace. Whenever it is invoked, it also receives a reference to the latest state in the current (counterfactual) trace. In that way, it can compare that state with the corresponding one in the actual trace and decide if intervention is necessary. If an infection happens in the counterfactual trace that did not happen in the actual trace, then the infection is suppressed as defined in the intervention rules. The intervention functor can be integrated very easily into Mesa's update loop. We simply pass a reference to the intervention functor to the main model class. In the `step` function (which updates the state of the entire model), we check for the existence of the intervention functor and, if present, invoke it by passing a reference to the current state. The intervention then takes place within the functor as described above and the simulation continues with the modified state.

Using the intervention object, we can now perform a causal analysis similar to the one discussed in Section 1. The goal is to determine the root cause (i.e. the agent that initiated the chain of infection) for each eventually infected agent. For ease of explanation, we restrict the population to 10 agents and simulation execution to 100 ticks. In the beginning of the simulation, two agents (agent 0 and agent 1) are infected, all other agents start off healthy. In each time step, one randomly chosen agent is updated according to the following rule. If the chosen agent is infected, it picks a neighbour n randomly and, if n is healthy, infects it with a probability of 50%. Furthermore, an infected agent has a 5% chance of recovering[11]. In order to be able to verify the correctness of the analysis results described further below, we record in the simulation the chain of infections. The evolution of one particular simulation run is shown in Table 6a. We can see that, after 100 ticks, all 10 agents are infected. By following the chain, we can, for example, infer that agent 6 was originally infected by agent 0 and agent 7 was originally infected by agent 1.

In order to find the root causes for each of the 10 eventually infected agents, we perform a counterfactual experiment by manipulating the initial infection states and replaying the simulation[12]. We start with a simple counterfactual experiment *without* intervention. As shown in Tables 6b and 6c, this approach is susceptible to preemption effects and thus not able to determine the root causes correctly. Table 6b shows the evolution of a counterfactual experiment without interventions in which agent 0's initial infection is deactivated. We can see that, despite the change, eventually all agents are infected. The same is the case in the experiment in which agent 1's initial

[10] The source code can be found on GitHub: https://github.com/bherd/pyCausalAnalysis

[11] The probabilities make no particular sense, they have just been chosen to allow for a significant number of agents to become infected eventually.

[12] In order to replay the original run precisely, it is important to eliminate randomness. This can be achieved technically by using in the counterfactual run the same random seed as in the original run.

Tick	Infections
17	1 → 3
26	0 → 9
33	3 → 2
37	0 → 4
46	0 → 5
55	4 → 6
66	2 → 7
79	1 → 8
evtl. inf.	{0,1,2,3,4, 5,6,7,8,9}

(a)

Tick	Infections
17	1 → 3
33	3 → 2
48	2 → 0
51	0 → 4
55	4 → 6
60	0 → 9
62	4 → 5
66	2 → 7
79	1 → 8
evtl. inf.	{0,1,2,3,4, 5,6,7,8,9}

(b)

Tick	Infections
26	0 → 9
37	0 → 4
46	0 → 5
55	4 → 6
65	0 → 3
70	4 → 1
72	3 → 7
79	1 → 8
88	1 → 2
evtl. inf.	{0,1,2,3,4, 5,6,7,8,9}

(c)

Tick	Infections
17	1 → 3
33	3 → 2
48	~~2 → 0~~
61	~~3 → 9~~
66	2 → 7
69	~~1 → 2~~
73	~~2 → 4~~
76	~~7 → 5~~
79	1 → 8
80	~~1 → 9~~
99	~~8 → 9~~
evtl. inf.	{1,2,3,7,8}

(d)

Tick	Infections
26	0 → 9
37	0 → 4
46	0 → 5
55	4 → 6
65	~~0 → 3~~
70	~~4 → 1~~
83	~~4 → 1~~
evtl. inf.	{0,4,5,6,9}

(e)

Fig. 6. Causal analysis of the virus propagation example. Base line run (a), counterfactual experiment **without intervention** with agent 0's infection deactivated (b) and with agent 1's infection deactivated (c), and counterfactual analysis **with intervention** with agent 0's infection deactivated (d) and with agent 1's infection deactivated (e)

infection is deactivated, as shown in Table 6c. The results illustrate nicely that a simple counterfactual analysis cannot discern true causal relationships from spurious ones.

If we perform the same experiment *with* interventions, the situation is significantly different. Table 6d shows the counterfactual experiment with intervention in which agent 0's initial infection is deactivated. We see that, during the simulation, a number of interventions take place (denoted by the crossed-out infection). Note that, as briefly mentioned in Section 4.4, the interventions also concern events that also happened in the actual trace but at a different time step. For example, in the original run, the infection of agent 9 happens at time 26; in the counterfactual run, it happens at time 61. Despite both events representing the infection of the same agent, they differ in their time steps and the infection in the counterfactual trace thus needs to be deactivated. After the simulation has finished, only agents 1,2,3,4, and 8 are infected. Table 6e shows the experiment in which agent 1's initial infection is deactivated. We see that the set of infected agents is different from the previous experiment. Both sets of eventually infected agents are disjoint and their union equals the total set of eventually infected agents. The analysis has thus produced a correct partitioning of infected agents into their respective root causes, concluding that agent 0 was the root cause for agents 4,5,6,9, and agent 1 was the root cause for agents 2,3,7,8. The correctness of the result can be verified by following the chain of infections in Table 6a.

6 RELATED WORK

Using a simulation model to detect general causal relationships between events (i.e. instances of type causation) is an integral part of almost any simulation experiment and generally done through repeated execution of the simulation and statistical analysis of the output data. This process can be further improved (e.g. with respect to spurious relationships) by extending it with Pearl's structural equation modelling as, for example, done by Kvassay *et al.* [14]. Their approach to the causal analysis of an agent-based simulation model consists of four steps: (i) *model analysis* utilising structural equation models and the idea of partitioning the causal factors according to their contribution, (ii) *implementation and data provisioning* in order to compute and log relevant causal analytical variables, (iii) *analysis* of the resulting simulation trace data, and (iv) *hypothesis validation* through appropriate calibration and manipulation of the simulation model. The approach shows promising results but it requires the modeller to have detailed knowledge about the mechanisms built into the model — an assumption which we explicitly exclude in our work.

No knowledge about the internal workings of the program under test is required by the explanation procedure proposed by Beer *et al.* [2]. The approach analyses an individual counterexample trace produced by a model checker with respect to its satisfaction of a linear temporal logic (LTL) formula. The basic idea is as follows. Given a counterexample trace π and an LTL formula ϕ such that ϕ fails on π, is there an event e in π that can be made *critical* for the satisfaction of ϕ? In other words, is it possible to set all values in the trace such that the satisfaction of ϕ becomes counterfactually dependent on e? If such an event exists, then it will qualify as a cause of ϕ's failure. Computing the set of causes for the failure of a given LTL formula is, in general, NP-complete but Beer *et al.* propose an algorithm that over-approximates the set in linear time. The approach is similar to ours in that it focusses on cases of token causation in traces. However, their focus on the *a-posteriori* detection of causes for the failure of an LTL formula (rather than on the detection of causal relationships between active events) results in different types of reported causes. Consider the scenario shown in Figure 4c and a hypothetical situation in which E does not happen. In searching for the set of causes, Beer's approach would include C (for reasons described in Section 4.5, paragraph 'Example 3') whereas our approach would explicitly consider it as a spurious relationship and exclude it.

7 CONCLUSIONS

Understanding the causal relationships within (possibly complex) simulation models is important for a number of reasons. First, it is a necessary requirement for *verification*, i.e. for assessing the correctness of the mechanisms built into the model. Second, it is important for *validation*, i.e. when assessing the accuracy with which a given model represents a certain real-world phenomenon. Third, it is crucial during *experimentation* when what–if scenarios are to be explored and true causal relationships are to be discerned from merely spurious ones. Historically, computational simulation models have been mostly treated as black boxes and causal analysis is reduced to controlled experimentation and statistical analysis of input–output relationships. As such, it is susceptible to the infamous 'correlation-versus-causation' problem. We argue in this paper that a purely observation-based process does not do justice to the potential of computer simulation models. We propose a formally grounded counterfactual analysis methodology for simulation-based causal analysis that is tailored to token causation. Given a simulation trace containing the stipulated cause and effect, it is based on the idea of interleaving simulation and manual intervention and controlling for randomness such that a counterfactual trace that is 'just similar enough' to the actual simulation trace and thus relevant for counterfactual analysis is generated. In that way, true causal relationships can be distinguished from mere spurious ones that result from irrelevant background effects.

Whilst the methodology described in this paper focusses on token causation, it can be easily generalised towards a probabilistic methodology for type causation as follows. Given a set T of simulation traces, we can say event A causes event B on a type level with probability p if the number of traces in which A acts as a token cause of B (according to the methodology of this paper) divided by the overall number of traces in T is equal to p. It is easy to see that this approach would be able to address the four problematic points described in Section 2 adequately, as argued below.

Imperfect regularities: The probabilistic approach is naturally capable of coping with imperfect regularities since it does not, by definition, assume that there is a strict regularity between causes and effects.

Irrelevance: If A is not relevant for B, our intervention rules ensure that manually turning off A would have no effect on the occurrence of B. A would thus not be identified as a cause of B.

Asymmetry: The sequential updating approach assumes causes to happen temporally before their effects, so asymmetry is inherently being dealt with.

Spurious regularities: Similar to the case of irrelevance, the intervention rules ensure that A is not reported as a cause of B if B is not counterfactually dependent upon A but A and B instead share a common cause (see example in Figure 4a).

At the current stage, the methodology is restricted to the detection of *basic* causal relationships, i.e. causal relationships between *active* events, and explicitly excludes specialised notions such as causation by omission and causation by prevention. As part of our future work, we aim to better understand the necessity for treating those notions and investigate whether different notions of causality can be detected using a uniform parametrisable approach. Another important point that we aim to investigate is to what extent the described approach can be applied to cases of causation between different ontological levels in complex adaptive systems, e.g. downward causation [1]. In that case, the events acting as causes are themselves emergent and non-atomic in nature. Furthermore, we believe (as indicated above) that a robust approach to token causation can be easily generalised to the type level. We aim to investigate this idea further by developing the ideas described in this paper into the direction of a larger explanatory framework. Finally, we aim to apply the methodology to a set of comprehensive real-world simulation models in which complex and non-obvious relationships exist.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] M. A. Bedau. 2002. Downward Causation and the Autonomy of Weak Emergence. Principia 6, 1 (2002), 5–50.

[2] I. Beer, S. Ben-David, H. Chockler, A. Orni, and R. Trefler. 2009. Explaining Counterexamples Using Causality. In Computer Aided Verification, Ahmed Bouajjani and Oded Maler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 94–108.

[3] T. Bosse and N. Mogles. 2013. Comparing Modelling Approaches in Aviation Safety. In Proceedings of the 4th International Air Transport and Operations Symposium, Toulouse, France, R. Curran (Ed.).

[4] S. Bouarfa, H. Blom, R. Curran, and M. Everdij. 2013. Agent-based modeling and simulation of emergent behavior in air transportation. Complex Adaptive Systems Modeling 1, 1 (2013), 1–26.

[5] T. Crane. 1998. Causality. In Philosophy: A Guide Through the Subject, A. C. Grayling (Ed.). Oxford University Press.

[6] E. Eells. 1991. Probabilistic Causality. Cambridge University Press.

[7] M. Guerini and A. Moneta. 2017. A method for agent-based models validation. Journal of Economic Dynamics and Control 82 (2017), 125 – 141.

[8] C. Hitchcock. 2001. The Intransitivity of Causation Revealed in Equations and Graphs. The Journal of Philosophy 98, 6 (2001), 273–299.

[9] C. Hitchcock. 2007. Prevention, preemption, and the principle of sufficient reason. The Philosophical Review 116, 4 (Oct 2007), 495–532.

[10] C. Hitchcock. 2012. Probabilistic Causation. In The Stanford Encyclopedia of Philosophy (winter 201 ed.), Edward N Zalta (Ed.).

[11] D. Hume. 2003. A Treatise of Human Nature. Courier Dover Publications.

[12] D. Kahneman and D. T. Miller. 1986. Norm theory: comparing reality to its alternatives. Psychological Review 93, 2 (1986), 136–153.

[13] S. Kleinberg and B. Mishra. 2009. The temporal logic of causal structures. In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 303–312.

[14] M. Kvassay, P. Krammer, L. Hluchý, and B. Schneider. 2017. Causal Analysis of an Agent-Based Model of Human Behaviour. Complexity 2017 (2017).

[15] R. Leombruni and M. Richiardi. 2005. Why are economists sceptical about agent-based simulations? Physica A: statistical Mechanics and Its Applications 355 (2005), 103–109.

[16] D. K. Lewis. 1986. On the Plurality of Worlds. Blackwell Publishers.

[17] C. M. Macal and M. J. North. 2010. Tutorial on agent-based modelling and simulation. Journal of Simulation 4, 3 (2010), 151–162.

[18] T. Maudlin. 2007. A Modest Proposal Concerning Laws, Counterfactuals, and Explanations. In The Metaphysics Within Physics. Oxford University Press, 5–49.

[19] R. McCune and G. Madey. 2013. Agent-based simulation of cooperative hunting with UAVs. In Proceedings of the Agent-Directed Simulation Symposium. Society for Computer Simulation International.

[20] S. McGrath. 2005. Causation by Omission: A Dilemma. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 123, 1/2 (2005), 125–148.

[21] L.A. Paul and Ned Hall. 2013. Causation: a user's guide. Oxford.

[22] J. Pearl. 2000. Causality: Models, Reasoning and Inference. Cambridge University Press.

[23] J. Schaffer. 2008. The Metaphysics of Causation. In The Stanford Encyclopedia of Philosophy (fall 2008 ed.), Edward N. Zalta (Ed.).

[24] I. Shpitser and J. Pearl. 2007. What Counterfactuals Can Be Tested. Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (2007), 352–359.

[25] J. M. Spivey. 1992. The Z notation: A Reference Manual. Prentice Hall International (UK) Ltd., Hertfordshire, UK.

[26] P. Suppes. 1970. A Probabilistic Theory of Causality. North-Holland Publishing Company, Amsterdam.

[27] A. Waldherr and N. Wijermans. 2013. Communicating Social Simulation Models to Sceptical Minds. Journal of Artificial Societies and Social Simulation 16, 4 (2013), 13.

[28] Y. Wei, G. Madey, and M. Blake. 2013. Agent-based simulation for UAV swarm mission planning and execution. In Proceedings of the Agent-Directed Simulation Symposium. Society for Computer Simulation International, 2.

[29] J. E. Wolff. 2016. Using Defaults to Understand Token Causation. Journal of Philosophy 113, 1 (2016), 5–26.

[30] S. Yablo. 2002. De Facto Dependence. The Journal of Philosophy 99, 3 (2002), 130–148.