What caused what? An irreducible account of actual causation.
Larissa Albantakis\textsuperscript{1,}, William Marshall\textsuperscript{1}, Erik Hoel\textsuperscript{2}, Giulio Tononi\textsuperscript{1,3}

\textsuperscript{1}Department of Psychiatry, Wisconsin Institute for Sleep and Consciousness, University of Wisconsin-Madison, WI, USA
\textsuperscript{2}Department of Biological Sciences, Columbia University, New York, NY, USA
\textsuperscript{*}Corresponding authors: albantakis@wisc.edu, gtononi@wisc.edu

Actual causation is concerned with the question “what caused what?”. Consider a transition between two subsequent observations within a system of elements. Even under perfect knowledge of the system, a straightforward answer to this question may not be available. Counterfactual accounts of actual causation based on graphical models, paired with system interventions, have demonstrated initial success in addressing specific problem cases. We present a formal account of actual causation, applicable to discrete dynamical systems of interacting elements, that considers all counterfactual states of a state transition from \( t - 1 \) to \( t \). Within such a transition, causal links are considered from two complementary points of view: we can ask if any occurrence at time \( t \) has an actual cause at \( t - 1 \), but also if any occurrence at time \( t - 1 \) has an actual effect at \( t \). We address the problem of identifying such actual causes and actual effects in a principled manner by starting from a set of basic requirements for causation (existence, composition, information, integration, and exclusion). We present a formal framework to implement these requirements based on system manipulations and partitions. This framework is used to provide a complete causal account of the transition by identifying and quantifying the strength of all actual causes and effects linking two occurrences. Finally, we examine several exemplary cases and paradoxes of causation and show that they can be illuminated by the proposed framework for quantifying actual causation.

1. Introduction

The nature of cause and effect has been much debated in both philosophy and the sciences. To date, there is no single widely accepted account of causation, and the various sciences focus on different aspects of the issue (Illari et al., 2011). In physics, no formal notion of causation seems even required to describe the dynamical evolution of a physical system by a set of mathematical laws. At most, the notion of causation is reduced to the most basic requirement that causes must precede and be able to influence their effects. Accordingly, physics restricts causes to occurrences within the past light cone of their effects—no further constraints are imposed as to “what caused what”. However, moving beyond fundamental physics and its reliance on laws of nature, the search for causal mechanisms behind observed phenomena is an important goal of scientific inquiry.

In this context, two broad aspects of causation can be distinguished: Natural, health, and social sciences are typically concerned with general (or type) causation—the question whether the type of occurrence A generally “brings about” the type of occurrence B, e.g., whether stress causes illness. By contrast, computer science, philosophy, and the law, tend to focus on the underlying notion of actual (or token) causation—the question “what caused what” given a specific occurrence A followed by a specific occurrence B. For example, Anne was hurrying to an important appointment on a slippery road, wearing high-heels; she fell and broke her leg. Did the fact that she was wearing high-heels cause her injury, or was it the slippery road, her running, or any combination of the above? Even with perfect knowledge of all circumstances, the prior system state, and the outcome, there often is no straightforward answer to the “what caused what” question. This has been demonstrated by a long list of controversial examples conceived, analyzed, and debated primarily by philosophers (Lewis, 1986; Pearl, 2000; Paul and Hall, 2013; Halpern, 2016).

During the last decades, a number of attempts to operationalize the notion of causation and to give it a formal description have been developed, most notably in computer science, probability theory, statistics (Good, 1961; Suppes, 1970; Pearl, 1988, 2000), and neuroscience, e.g., (Tononi et al., 1999). Important advances have been made in providing criteria for distinguishing causal relations from mere correlations. This distinction is important for controlling a system via interventions, since in general it is not possible to predict how a system will react to perturbation purely from observed correlations\textsuperscript{1}. In other words, merely observing that occurrence A is reliably followed by occurrence B is insufficient to establish whether the relation between A and B is causal or not. Leaving aside metaphysical accounts of the nature of causation (see Paul and Hall, 2013, Chapter 2), for an overview), formal frameworks of causal analysis invariably rely on the notion of counterfactuals (Lewis, 1973; Pearl, 2000; Woodward, 2004; Yablo, 2004)—alternative

\textsuperscript{1} For example, observing a positive correlation between the number of predator species A and B, both eating rabbits, does not imply that species A will decrease if species B is diminished by intervention. The opposite may be the case (because now there is more rabbits for species A).
occurrences “counter to fact”. Graphical methods paired with system interventions (Pearl, 2000) have proven especially valuable for developing causal explanations. Given a causal network that represents how the state of each variable depends on other system variables via a “structural equation” (Pearl, 2000), the effects of interventions imposed from outside the network can be evaluated by setting certain variables to a specific value. This operation has been formalized by Pearl, who introduced the “do-operator”, do($X=x$), which signifies that a subset of system variables $X$ has been actively set into state $x$ rather than being passively observed in this state (Pearl, 2000). In contrast to statistical, ’Bayesian’ networks, causal networks, are thus characterized through dependencies identified by intervention, not observation. Because statistical dependence does not imply causal dependence, and vice versa (Ay and Polani, 2008), the conditional probability of occurrence $B$ after observing occurrence $A$, $p(B|A)$ may differ from the probability of occurrence $B$ after enforcing $A$ $p(B|do(A))$.

The causal networks approach has also been applied to the case of *actual causation* (Pearl, 2000; Halpern and Pearl, 2005; Halpern, 2015), where statistical inferences about populations or processes rather than individual occurrences cannot be meaningfully applied. In this framework, system interventions can be used to evaluate whether and to what extent an occurrence was necessary or sufficient for a subsequent occurrence by assessing counterfactuals. Approaches differ primarily depending on which sets of variables are held fixed and which are allowed to vary during the analysis—in other words—which counterfactual states are taken into account. While promising results have been obtained in specific cases, no single proposal to date has characterized actual causation in a universally satisfying manner (Paul and Hall, 2013; Halpern, 2016).

Below we present a formal account of actual causation that considers *all* counterfactual states, which allows us to express causal analysis in probabilistic, informational terms. We aim at providing a causal account of “what caused what”, given a transition $X_{t-1} = x_{t-1} < Y_t = y_t$ (“$x_{t-1}$ precedes $y_t$”) between two subsequent observations within a discrete dynamical system $S$ with $X_{t-1}, Y_t \subseteq S$, constituted of interacting elements (Fig. 1). Within such a transition, unlike previous accounts of actual causation (e.g., Pearl, 2000; Paul and Hall, 2013; Halpern, 2016), but see (Chajewski and Halpern, 1997)), causal links are considered from the perspective of both causes and effects. Specifically, we ask both if an occurrence $Y_t = y_t \subseteq Y_t = y_t$ at time $t$ has an actual cause at $t-1$; and also, if an occurrence $X_{t-1} = x_{t-1} \subseteq X_{t-1} = x_{t-1}$ at time $t-1$ has an actual effect at $t$. We identify such actual causes and actual effects and demonstrate that both perspectives are essential to give a complete causal account of the transition $X_{t-1} = x_{t-1} < Y_t = y_t$. Moreover, by considering all possible counterfactual states, we quantify the strength of causal links between occurrences and their actual cause/effect in informational terms.

The analysis is based on five principles that characterize *potential causation*—the causal constraints exerted by a mechanism in a given state—namely intrinsic existence, composition, information, integration, and exclusion, identified in the context of integrated information theory (IIT; Oizumi et al., 2014; Albantakis and Tononi, 2015). Originally developed as a theory of consciousness (Tononi, 2015; Tononi et al., 2016), IIT provides the tools to identify all potential causes and effects within a discrete dynamical system in a state-dependent manner and to quantify their power. Here we employ this very set of tools to identify actual causes and effects and to quantify their strength. This formal framework for actual causation is based on first principles and is applicable to a wide range of systems, whether deterministic and probabilistic, with feedforward and recurrent architectures.

In the following, we will first formally describe the proposed causal analysis. We then demonstrate its utility on a set of examples, which illustrate the need to characterize both causes and effects, the fact that causation can be compositional, and the importance of identifying irreducible causes and effects for obtaining a complete causal account. Finally, we illustrate several prominent paradoxical cases from the actual causation literature, including overdetermination and prevention.

### 2. Theory

IIT’s approach to *potential* causation is concerned with the *intrinsic cause-effect power* of a physical system (*intrinsic existence*). The IIT formalism starts from a discrete dynamical system $S$ in its current state $s_{t}$ and asks how the system’s elements, alone and in combination (*composition*), constrain the potential past and future states of the system (*information*), and whether they do so above and beyond their parts (*integration*). Using the measures developed within IIT, for any subset $S = s_{t} \subseteq S = s_{t}$, we can identify the maximally irreducible set of potential causes and effects within the system (*exclusion*), and quantify its irreducible cause-effect power (*integrated information $q$*) (Oizumi et al., 2014; Tononi, 2015). Translated to the case of *actual* causation, IIT’s five principles are as follows:

1. **Existence:** Causes and effect are actual. The actual cause of an occurrence $Y_t = y_t$ must actually have happened at $t - 1$, and the actual effect of an occurrence $X_{t-1} = x_{t-1}$ at $t = 0$ has an actual cause at $t - 1$.

---

2 Note that counterfactuals here strictly refer to possible states within the system’s state space other than the actual one and not to abstract notions such as other “possible worlds” as in (Lewis, 1973), see also (Pearl, 2000, Chapter 7).
Actual causation \( X_{t-1} = x_{t-1} \) must actually have happened at \( t \), in a system that actually exists. An actual cause is thus an occurrence at \( t - 1 \) and an actual effect an occurrence at \( t \) as observed in that system.\(^3\)

(2) Composition: Causes and effects are structured. Any subset of an observation can be an occurrence with its own actual cause or effect. Likewise, a subset of an actual cause/effect can itself also be a separate actual cause/effect of a different occurrence within the transition.

(3) Information: Causes and effects are specific. An occurrence must increase the probability that its actual cause/effect occurred compared to its expected probability when the occurrence is not specified (that is, when all possible states of the occurrence are considered with equal probability). In other words, an actual cause/effect must be distinguishable from noise.

(4) Integration: Causes and effects are irreducible. Only irreducible occurrences can have actual causes or effects. An occurrence must determine its actual cause/effect irreducibly, above and beyond its parts.

(5) Exclusion: Causes and effects are definite. An occurrence can have at most one actual cause/effect, which is the smallest set of elements whose state is most irreducibly determined by the occurrence.

\( X_{t-1} = x_{t-1} \) must actually have happened at \( t \), in a system that actually exists. An actual cause is thus an occurrence at \( t - 1 \) and an actual effect an occurrence at \( t \) as observed in that system.\(^3\)

2.1 Actual causes and actual effects

The objective of this section is to introduce the notion of a causal account for a transition of interest \( x_{t-1} < y_t \) as the set of all causal links within the transition. A causal link exists between an occurrence \( x_{t-1} \subseteq x_{t-1} \) and its actual effect at \( t \), or an occurrence \( y_t \subseteq y_t \) and its actual cause at \( t - 1 \).

\(^3\) Note that this is also a requirement in the original and modified versions of the Halpern and Pearl account of actual causation (Halpern and Pearl, 2005; Halpern, 2015). It corresponds to their first clause ("AC1"), that for \( X = x_{t-1} \) to be an actual cause of \( Y = y_t \) both must actually happen in the first place.

To apply these principles to the analysis of actual causation, we assume a discrete dynamical system \( S \), constituted of \( n \) interacting elements \( S_i \) with \( i = 1, ..., n \) (Fig. 1A). Each element must have at least two internal states, which can be observed and manipulated, and is equipped with a Markovian input-output function \( f_i \) that determines the element’s output state depending only on the previous system state: \( s_{i,t} = f_i(s_{i,t-1} = s_{i,t-1}) \). This means that all elements are conditionally independent given the past state \( s_{t-1} \) of the system and any system transition \( S_{t-1} = s_{t-1} \Leftrightarrow S_t = s_t \) can be represented within a directed, acyclic, causal network (Fig. 1B) (Pearl, 2000; Ay and Polani, 2008; Halpern, 2015). Alternatively, \( S \) is fully described by its transition probabilities (Fig. 1C):

\[
p(S_t = s_t | S_{t-1} = s_{t-1}) = \prod_{i=1}^{n} p(S_{i,t} = s_{i,t} | S_{i,t-1} = s_{i,t-1}), \quad \forall s_{t}, s_{t-1}.
\]

In contrast to other accounts of (actual) causation, we explicitly interpret \( S \) as a physical system with interacting physical elements, as opposed to abstract variables, like, for example, a ‘desire for chocolate’. Examples of physical systems include a set of neurons in the brain or logic gates in a computer. In reality, \( S \) will always be an (approximate) causal model of a given physical system, constructed from interventionist experiments\(^4\). However, our objective is to formulate a quantitative account of actual causation, without confounding issues due to incomplete knowledge, such as estimation biases of probabilities from finite sampling, or latent variables. For this reason, we assume full knowledge about the physical system, that is, we assume a perfect causal model that is equivalent to it. For simplicity, without loss of generality, we restrict our analysis to systems of binary elements. As the final prerequisites, we define an observation as the state of a subset of elements within system \( S \) at a particular instant, e.g. \( x_{t-1} = x_{t-1} \), or \( Y_t = y_t \) with \( X_{t-1}, Y_t \subseteq S \). To simplify notation, observations are subsequently denoted by their state only, e.g., \( y_t \) instead of \( Y_t = y_t \).

\(^4\) The transition probabilities can, in principle, be determined, by perturbing the system into all possible states and observing the resulting transitions. Alternatively, the causal network can be constructed by experimentally identifying the input-output function of each element. Crucially, merely observing the system without experimental manipulation is insufficient to identify causal relationships in most situations.
Below, we will formally define causal link, actual cause, actual effect, and causal account following the five principles outlined above: existence, composition, information, integration, and exclusion.

Existence. A transition of interest $x_{t-1} < y_t$ for a causal account is defined by a pair $\{X_{t-1} = x_{t-1}; Y_t = y_t\}$ of subsequent observations within a system $S$ (note that $X$ and $Y$ can contain the same elements), consistent with the system's transition probabilities.

Elements in $S \setminus X$ at $t - 1$ are considered exogenous and are treated as background conditions, meaning that their states are held fixed during the causal analysis. Bold letters are used for the two observations $\{x_{t-1}; y_t\}$ that define the transition $x_{t-1} < y_t$, while $x_{t-1}$ or $y_t$ denote occurrences within the transition, i.e., subsets of $x_{t-1}$ or $y_t$. As a first example, we consider the transition $x_{t-1} < y_t$ shown in Fig. 2A of the physical system $S$ depicted in Fig. 1. Here, $x_{t-1} = y_t = \{\text{OR, AND = 10}\}$, consistent with the system's transition probabilities (Fig. 1C).

Composition. For a causal account of the transition $x_{t-1} < y_t$, we are interested in determining all causal links between occurrences $x_{t-1}$ and their actual effects, and occurrences $y_t$ and their actual causes. To this end, we consider the power set of $x_{t-1}$ as occurrences that could have actual effects and the power set of $y_t$ as occurrences that could have actual causes. Multi-element occurrences are termed “high-order” occurrences in the following. Given a particular occurrence $x_{t-1}$, we test the power set of $y_t$ to identify the actual effect of $x_{t-1}$. Given a particular occurrence $y_t$, we test the power set of $x_{t-1}$ to identify the actual cause of $y_t$.

In the example transition shown in Fig. 2A, causal links may thus exist between the occurrences $\{\text{OR = 1}\}$, $\{\text{AND = 0}\}$, and $\{\text{OR, AND = 10}\}$ at $t - 1$ and $t$, and any of their possible effects or causes $\{\text{OR = 1}\}$, $\{\text{AND = 0}\}$, and $\{\text{OR, AND = 10}\}$ at $t$ and $t - 1$, respectively.

Information. In order to identify a causal link between two occurrences $x_{t-1}$ and $y_t$, it is not enough to consider the actual states $x_{t-1}$ and $y_t$; other, hypothetical states (counterfactuals) must also be taken into account (Lewis, 1973; Pearl, 2000). Our objective here is to quantify the strength of causal links rigorously, beyond merely identifying them. By taking all possible counterfactuals into account (Eqn. 1), we can express causal strength utilizing interventional conditional probabilities and probability distributions (Tononi et al., 1999; Ay and Polani, 2008; Balduzzi and Tononi, 2008; Hoel et al., 2013; Oizumi et al., 2014). To this end, we assume a maximum entropy distribution over the possible states of $X_{t-1}$ throughout the analysis, i.e., $p(X_{t-1} = x) = |\Omega_x|^{-1}$ ∀$x \in \Omega_x$, where $\Omega_x$ denotes the set of all possible states of $X$ (see Discussion 4.1). This ensures that causal strength reflects constraints due to the occurrence itself, rather than a predisposition of elements to be in one state rather than another. Roughly speaking, we assess how much of a difference an occurrence $x_{t-1}$ (or $y_t$) makes to the probability of a candidate effect (or cause) compared to all possible states of $X_{t-1}$ (or $Y_t$).

To identify actual causes and effects and to quantify causal links, we start from the notions of a cause- and effect repertoire—the interventional conditional probability distributions over the potential past and future states of the system (Oizumi et al., 2014; Albantakis and Tononi, 2015), see also (Tononi, 2015; Marshall et al., 2016b) for a general mathematical definition. The effect repertoire specifies how an occurrence $x_{t-1}$ constrains the potential future states of a set of elements $Y_t$ (potential effects). Likewise, the cause repertoire specifies how an occurrence $y_t$ constrains the potential past states of a set of elements $X_{t-1}$ (potential causes) (Fig. 2B, C).

The effect repertoire of an occurrence $x_{t-1}$ over a single element $Y_{t,t}$ is a conditional probability distribution over $Y_t$ given $x_{t-1}$ in the causal network. The complementary cause repertoire of a single element
occurrence at time $t$, $Y_{t,t} = y_{t,t}$, is a conditional probability distribution over a set of elements $X_{t-1}$ given $Y_{t}$. These and all subsequently defined probabilities and probability distributions can be derived from the system’s transition probabilities (Eqn. 1) by conditioning on the state of the occurrence and marginalizing across the states of undetermined elements $X_{t-1} \backslash X_{t-1}$ and $Y_{t} \backslash Y_{t}$. Explicitly, for a single element occurrence $Y_{t,t}$ the effect repertoire is:

$$\pi(Y_{t,t} | x_{t-1}) = \frac{1}{|O_{X \backslash X}|} \sum_{x \in O_{X \backslash X}} p(Y_{t,t} | do(x_{t-1}, X_{t-1} \backslash X_{t-1} = z)),$$

and the corresponding cause repertoire, using Bayes’ rule, is:

$$\pi(X_{t-1} | Y_{t,t}) = \frac{\sum_{x \in O_{X \backslash X}} p(Y_{t,t} | do(x_{t-1}, X_{t-1} \backslash X_{t-1} = z))}{\sum_{x \in O_{X \backslash X}} p(Y_{t,t} | do(X = z))}.$$

Cause- and effect repertoires capture all the constraints that the occurrence $x_{t-1}$ exerts on a set of elements $Y_{t}$, and vice versa for $y_{t}$ and $X_{t-1}$. However, in the general case of multiple elements in $Y_{t}$, the conditional probability distribution $p(Y_{t,t} | x_{t-1})$ not only contains dependencies of $Y_{t}$ on $x_{t-1}$, but also correlations between elements in $Y_{t}$ due to common inputs from elements in $X_{t-1} \backslash X_{t-1}$, which should not be counted as constraints due to $x_{t-1}$. To discount such correlations, we define the effect repertoire over a set of elements $Y_{t}$ as:

$$\pi(Y_{t} | x_{t-1}) = \prod_{i} \pi(Y_{t,i} | x_{t-1}).$$

In the same manner, we define the cause repertoire as:

$$\pi(X_{t-1} | y_{t}) = \frac{1}{Q} \prod_{i} \pi(X_{t-1,i} | y_{t,i}).$$

with

$$Q = \sum_{x \in O_{X \backslash X}} \prod_{i} \pi(X_{t-1,i} = x | y_{t,i}).$$

In summary, the effect- and cause repertoires $\pi(Y_{t} | x_{t-1})$ and $\pi(X_{t-1} | y_{t})$, respectively, are interventional conditional probability distributions that specify constraints due to an occurrence on the potential past and future states of a set of system elements, enforcing a maximum entropy distribution over the possible states of $X_{t-1}$ and eliminating correlative constraints due to common inputs.

In the context of actual causation, the question of interest is “what caused what” given knowledge of “what actually happened” over two consecutive time steps. To this end, given the actual state of $Y_{t}$, we define an effect ratio $\rho_{e}$ for the occurrence $x_{t-1}$ and a subsequent occurrence $y_{t}$ within the transition $x_{t-1} \prec y_{t}$, as:

$$\rho_{e}(x_{t-1}, y_{t}) = \log_{2}(\pi(y_{t} | x_{t-1})/\pi(y_{t})), \quad (4)$$

where

$$\pi(y_{t}) = \prod_{i} \pi(y_{t,i}),$$

and

$$\pi(Y_{t,t} | x_{t-1} = x) = \frac{1}{|O_{X \backslash X}|} \sum_{x \in O_{X \backslash X}} p(Y_{t,t} | do(X_{t-1} = x)).$$

In words, the effect ratio $\rho_{e}$ is the probability of an occurrence at $t$ when constrained by an occurrence at $t - 1$, divided by its probability when unconstrained. The unconstrained effect repertoire $\pi(y_{t,t})$ is the probability of $Y_{t,t} = y_{t,t}$ given a uniform distribution of all possible input states to element $Y_{t,t}$.

Similarly, given the actual state of $X_{t-1}$, we define the cause ratio $\rho_{c}$ for the occurrence $y_{t}$ and a prior occurrence $x_{t-1}$ within the transition $x_{t-1} \prec y_{t}$, as:

$$\rho_{c}(x_{t-1}, y_{t}) = \log_{2}(\pi(x_{t-1} | y_{t})/\pi(x_{t-1})). \quad (5)$$

In words, the cause ratio $\rho_{c}$ is the probability of an occurrence at $t - 1$ when constrained by an occurrence at $t$, divided by its probability when unconstrained. The unconstrained cause repertoire $\pi(x_{t-1} | y_{t})$ is a uniform distribution over the possible states of $X_{t-1}$. Using base 2 logarithms in Eqns. 4 and 5 sets the unit of $\rho$ to bits.

A positive effect ratio $\rho_{e}(x_{t-1}, y_{t}) > 0$ means that the occurrence $x_{t-1}$ increases the likelihood of the particular occurrence $y_{t}$, compared to its probability in the case of maximal uncertainty about the state of $X_{t-1}$. The occurrence $x_{t-1}$ thus makes a positive difference in bringing about $y_{t}$. This is a necessary, but not sufficient condition for $y_{t}$ to be an actual effect of $x_{t-1}$. Similarly, a positive cause ratio $\rho_{c}(x_{t-1}, y_{t}) > 0$ means that the occurrence $y_{t}$ increases the likelihood of the particular occurrence $x_{t-1}$ having happened, compared to maximum entropy $|O_{X \backslash X}|^{-1}$. Again, this is a necessary, but not sufficient condition for $x_{t-1}$ to be an actual cause of $y_{t}$. Both $\rho_{e}$ and $\rho_{c}$ can be interpreted as measures of the information that one occurrence specifies about the other.

In Fig. 3A, for example, the occurrence $x_{t-1} = \{\text{OR} = 1\}$ raises the probability of $y_{t} = \{\text{OR} = 1\}$, and vice versa (Fig. 3B), with $\rho_{e}(x_{t-1}, y_{t}) = \rho_{c}(x_{t-1}, y_{t}) = 0.415$ bits.\footnote{In general, $\pi(Y_{t} | x_{t-1}) = p(Y_{t} | x_{t-1})$. It is only in the special case that $X_{t-1}$ already includes all inputs (all parents) of $Y_{t}$, or determines $Y_{t}$ completely, that $\pi(Y_{t} | x_{t-1})$ is equivalent to $p(Y_{t} | x_{t-1})$ because of the conditional independence of all $Y_{t}$ (see Eqn. 1).}

By contrast, the occurrence $x_{t-1} = \{\text{OR} = 1\}$ lowers the probability of occurrence $y_{t} = \{\text{AND} = 0\}$ and also of the 2nd-order occurrence $y_{t} = \{\text{OR, AND} = 10\}$. Thus, neither $y_{t} = \{\text{AND} = 0\}$ nor $y_{t} = \{\text{OR, AND} = 10\}$ can be actual effects of $x_{t-1} = \{\text{OR} = 1\}$. Likewise, the occurrence $y_{t} = \{\text{OR} = 1\}$ lowers the probability of $x_{t-1} = \{\text{AND} = 0\}$, which can thus not be the actual cause of $y_{t}$.

In general, the cause and effect ratios for an occurrence pair $(x_{t-1}, y_{t})$ are not identical. In other words, there is no equivalent of Bayes’ rule for $\pi$, as generally $\pi(y_{t}) \pi(x_{t-1} | y_{t}) \neq \pi(x_{t-1}) \pi(y_{t} | x_{t-1})$, since Eqn 2 and 3 do not follow the standard definitions for conditional probability distributions.
Fig. 3: Information and Integration. (A, B) Example effect and cause ratios. The state that actually occurred is selected from the effect or cause repertoire (green is used for effects, blue for causes). Its probability is compared to the probability of the same state without constraints (overlaid distributions without fill). (C, D) Irreducible effect and cause ratios. The probability of the actual state in the effect or cause repertoire is compared against its probability in the partitioned effect or cause repertoire (overlaid distributions without fill). Of all 2^{nd}-order occurrences shown, only $y_t = \{\text{OR, AND} = 10\}$ irreducibly constrains $x_{t-1} = \{\text{OR, AND} = 10\}$.

**Integration.** The irreducibility of a candidate causal link is tested by “disintegrating” it into independent parts, that is, by rendering the connections between the parts ineffective (Balduzzi and Tononi, 2008; Janzing et al., 2013; Oizumi et al., 2014; Albantakis and Tononi, 2015). Practically, this is accomplished by taking the product of the marginal distributions corresponding to each part.

In the case of a candidate causal link from $x_{t-1}$ to $y_t$ ($x_{t-1} \rightarrow y_t$), the effect repertoire under the disintegrating partition $P$ is defined as:

$$\pi(y_t | x_{t-1})_P = \prod_{p \in \{1, \ldots, m\}} \pi(y_p | x_{p,t-1}) \times \pi(y_t \setminus \cup_{p \in \{1, \ldots, m\}} y_{p,t}),$$

(6)

where the occurrence $x_{t-1}$ is partitioned into $1 \leq m \leq |x_{t-1}|$ parts $x_{p,t-1}, p = 1 \ldots m$. Each of these $x_{p,t-1}$ is then paired with a different, potentially empty, non-overlapping set $Y_{p,t} \subseteq Y_t$, with the additional condition that if $m = 1$ then $Y_{1,t} = \emptyset$. The product of the corresponding $m$ effect repertoires is multiplied by the unconditional effect repertoire of the remaining elements $y_t \setminus \cup_{p \in \{1, \ldots, m\}} y_{p,t}$. In case there are no such left-over elements of $Y_t$, $\pi(y_t \setminus \cup_{p \in \{1, \ldots, m\}} y_{p,t}) = \pi(\emptyset) = 1$. Defined in this way, the disintegrating partition $P$ guarantees that the partitioned occurrence could not have an irreducible effect on any subset of $Y_t$, as any constraints that $x_{t-1}$ might have on $Y_t$ are split into independent parts.

In the same way, the cause repertoire under a disintegrating partition $P$ of the candidate causal link from $y_t$ to $x_{t-1}$ ($x_{t-1} \leftarrow y_t$) is defined as:

$$\pi(x_{t-1} | y_t)_P = \prod_{p \in \{1, \ldots, m\}} \pi(x_{p,t-1} | y_{p,t}) \times \pi(x_{t-1} \setminus \cup_{p \in \{1, \ldots, m\}} x_{p,t-1}),$$

(7)

where now the occurrence $y_t$ is partitioned into $1 \leq m \leq |y_t|$ parts $y_{p,t}, p = 1 \ldots m$. Each of these $y_{p,t}$ is then paired with a different, potentially empty, non-overlapping set $X_{p,t-1} \subseteq X_{t-1}$, with $X_{1,t-1} = \emptyset$ if $m = 1$. The product of the corresponding $m$ cause repertoires is multiplied by the unconditional cause repertoire of the remaining elements $X_{t-1} \setminus \cup_{p \in \{1, \ldots, m\}} X_{p,t-1}$, again with $\pi(\emptyset) = 1$. Defined in this way, the disintegrating partition $P$ guarantees that the partitioned occurrence could not have an irreducible cause in any subset of $X_{t-1}$, as any constraints that $y_t$ might have on $X_{t-1}$ are split into independent parts.
Since single element occurrences cannot be further partitioned, \( m \) has to be 1. In that case, only a complete partition between \( x_{t-1} \) and \( y_t \) disintegrates the occurrence and Eqn. 6 and 7 thus reduce to \( \pi(y_t) \) and \( \pi(x_{t-1}) \), respectively.

By comparing \( \rho_e(x_{t-1}, y_t) \) against \( \rho_e(x_{t-1}, y_t)_P \), we can measure the difference a partition \( P \) makes to the effect ratio of the occurrence \( x_{t-1} \) on a candidate effect \( y_t \), and similarly for \( \rho_e(x_{t-1}, y_t) \). However, there are many ways to partition an occurrence: we define \( P(x_{t-1} \rightarrow y_t) \) as the set of all possible disintegrating partitions of an occurrence \( x_{t-1} \) and its candidate effect \( y_t \), and \( P(x_{t-1} \leftarrow y_t) \) as the set of all possible disintegrating partitions of an occurrence \( y_t \) and its candidate cause \( x_{t-1} \). A measure of irreducibility requires choosing the partition \( P \) that makes the least difference to the ratio, which is denoted the minimum information partition (MIP) (Oizumi et al., 2014; Albantakis and Tononi, 2015), respectively:

\[
\text{MIP}(x_{t-1} \rightarrow y_t) = \arg\min_P \rho_e(x_{t-1}, y_t) - \rho_e(x_{t-1}, y_t)_P
\]

and

\[
\text{MIP}(x_{t-1} \leftarrow y_t) = \arg\min_P \rho_e(x_{t-1}, y_t) - \rho_e(x_{t-1}, y_t)_P.
\]

Note that the MIP will typically depend on the specific transition \( x_{t-1} \rightarrow y_t \), however we omit this dependence from notation for simplicity, as it will be clear from the context which transition is being considered.

Taken together, we define the irreducible effect ratio \( \alpha_e \) as the difference between the intact ratio and the ratio under the MIP:

\[
\alpha_e(x_{t-1}, y_t) = \rho_e(x_{t-1}, y_t) - \rho_e(x_{t-1}, y_t)_{\text{MIP}} = \log_2(\pi(y_t | x_{t-1}) / \pi(y_t | x_{t-1})_{\text{MIP}}).
\]

The irreducible cause ratio \( \alpha_c \):

\[
\alpha_c(x_{t-1}, y_t) = \rho_e(x_{t-1}, y_t) - \rho_e(x_{t-1}, y_t)_{\text{MIP}} = \log_2(\pi(x_{t-1} | y_t) / \pi(x_{t-1} | y_t)_{\text{MIP}}).
\]

A positive irreducible effect ratio (\( \alpha_e(x_{t-1}, y_t) > 0 \)) signifies that the occurrence \( x_{t-1} \) has an irreducible effect on \( y_t \), which is necessary but not sufficient for \( y_t \) to be an actual effect of \( x_{t-1} \). Likewise, a positive irreducible cause ratio (\( \alpha_c(x_{t-1}, y_t) > 0 \)) means that \( y_t \) has an irreducible cause in \( x_{t-1} \), which is a necessary but not sufficient condition for \( x_{t-1} \) to be an actual cause of \( y_t \). Note that for single element occurrences \( \alpha(x_{t-1}, y_t) = \rho(x_{t-1}, y_t) \), because only the complete partition disintegrates the occurrence.

In our example transition, the occurrence \( x_{t-1} = \{\text{OR, AND} = 10\} \) (Fig. 3C) is reducible. This is because \( x_{t-1} = \{\text{OR} = 1\} \) is sufficient to determine the occurrence \( y_t = \{\text{OR} = 1\} \) with probability 1.0 and \( x_{t-1} = \{\text{AND} = 0\} \) is sufficient to determine the occurrence \( y_t = \{\text{AND} = 0\} \) with probability 1.0, thus there is nothing to be gained by considering the two elements together as a 2nd-order occurrence at \( t - 1 \). By contrast, the occurrence \( y_t = \{\text{OR, AND} = 10\} \) determines the particular past state of \( x_{t-1} = \{\text{OR, AND} = 10\} \) with higher probability than the two single occurrences \( y_t = \{\text{OR} = 1\} \) and \( y_t = \{\text{AND} = 0\} \) taken separately (Fig. 3D, right). Thus, the 2nd-order occurrence \( y_t = \{\text{OR, AND} = 10\} \) is irreducible (\( \alpha_e(x_{t-1}, y_t) = 0.17 \) bits, with \( x_{t-1} = \{\text{OR, AND} = 10\} \)) (see Discussion 4.4).

**Exclusion:** For a particular occurrence \( x_{t-1} \), the power set of \( y_t \) has to be considered as candidate effects, and the power set of \( x_{t-1} \) has to be considered as candidate causes of a particular occurrence \( y_t \). It is possible that an occurrence \( x_{t-1} \) has multiple candidate effects \( y_t \) for which \( \alpha_e(x_{t-1}, y_t) > 0 \), as well as that an occurrence \( y_t \) has multiple candidate causes \( x_{t-1} \) for which \( \alpha_e(x_{t-1}, y_t) > 0 \). We define the irreducibility of an occurrence as its maximum irreducible effect (or cause) ratio over all possible candidate effects (or causes),

\[
\alpha_e^{\text{max}}(x_{t-1}) = \max_{y_t \subseteq y_t} \alpha_e(x_{t-1}, y_t),
\]

and

\[
\alpha_c^{\text{max}}(y_t) = \max_{x_{t-1} \subseteq x_{t-1}} \alpha_c(x_{t-1}, y_t).
\]

If \( \alpha^{\text{max}} = 0 \), then the occurrence is said to be reducible.

The exclusion principle enforces that every occurrence can at most have one actual cause and one actual effect, which is the smallest set of elements whose state is most irreducibly determined by the occurrence.7 For example, \( y_t = \{\text{OR} = 1\} \) has two candidate causes \( x_{t-1} = \{\text{OR} = 1\} \) and \( x_{t-1} = \{\text{OR, AND} = 10\} \) with \( \alpha_e(x_{t-1}, y_t) = \alpha_e^{\text{max}}(y_t) = 0.415 \) bits. In this case, \( x_{t-1} = \{\text{OR} = 1\} \) is the actual cause of \( y_t = \{\text{OR} = 1\} \). The exclusion principle thus avoids causal overdetermination, which arises from counting multiple causes or effects for a single occurrence. Note, however, that in a perfectly symmetric system indeterminism about the actual cause or effect is possible (see Results).

Having established the above causal principles, we are now in a position to formally define the actual cause and the actual effect of an occurrence:

**Definition 2.1.a** The actual cause of an irreducible occurrence \( y_t \subseteq y_t \) within the transition \( x_{t-1} \rightarrow y_t \) is the minimal set of elements \( x_{t-1} \) for which the irreducible cause ratio of \( y_t \) is maximal:

\[
x_{t-1}^{\ast}(y_t) = \arg\min_{x \subseteq y_t} |x|,
\]

7 The exclusion requirement is related to the third clause (“AC3”) in the original and modified versions of actual causation as proposed by (Halpern and Pearl, 2005) and (Halpern, 2015), which states that the actual cause must be minimal, meaning that no subset can also satisfy the conditions for being an actual cause.
$x'(y_t) = \{x_{t-1} \subseteq x_{t-1}[\alpha_e(x_{t-1}, y_t) = \alpha_e^{\max}(y_t)]\}$.

**Definition 2.1.b:** The *actual effect* of an irreducible occurrence $x_{t-1} \subseteq x_{t-1}$ within the transition $x_{t-1} \prec y_t$ is the minimal set of elements $y'_t$ for which the irreducible effect ratio of $x_{t-1}$ is maximal:

$y'_t(x_{t-1}) = \arg\min_{y \in y'(x_{t-1})} |y|$, where

$y'(x_{t-1}) = \{y_t \subseteq y_t | \alpha_e(x_{t-1}, y_t) = \alpha_e^{\max}(x_{t-1})\}$.

Based on definitions 2.1.a) and b):

**Definition 2.2:** A *causal link* within a transition $x_{t-1} \prec y_t$ consists of an occurrence $x_{t-1} \subseteq x_{t-1}$ with an actual effect $y'_t(x_{t-1})$ and its associated maximally irreducible effect ratio $\alpha_e^{\max}(x_{t-1}) > 0$, or an occurrence $y_t \subseteq y_t$ with an actual cause $x'_{t-1}(y_t)$ and its associated maximally irreducible cause ratio $\alpha_e^{\max}(y_t) > 0$:

$x_{t-1} \rightarrow y_t = \{x_{t-1}, y_t = y'_t(x_t), \alpha_e^{\max}(x_{t-1})\}$

or

$x_{t-1} \leftarrow y_t = \{y_t, x_{t-1} = x'_{t-1}(y_t), \alpha_e^{\max}(y_t)\}$

The strength of a causal link is determined by its $\alpha_e^{\max}$ or $\alpha_e^{\max}$ value. Reducible occurrences ($\alpha_e^{\max}(x_{t-1}) = 0$ or $\alpha_e^{\max}(y_t) = 0$) cannot form a causal link.

**Definition 2.3:** A *causal account* $C$ of a transition $x_{t-1} \prec y_t$ is the set of all causal links $x_{t-1} \rightarrow y_t$ and $x_{t-1} \leftarrow y_t$ within the transition.

Under this definition, all actual causes and actual effects contribute to the causal account $C(x_{t-1} \prec y_t)$. Crucially, the fact that there is a causal link $x_{t-1} \rightarrow y_t$ does not necessarily imply that the reverse causal link $x_{t-1} \leftarrow y_t$ exists, and vice versa. In other words, just because $y_t$ is the actual effect of $x_{t-1}$, the occurrence $x_{t-1}$ does not have to be the actual cause of $y_t$. It is therefore not redundant to include both directions in $C(x_{t-1} \prec y_t)$, as illustrated by examples of overdetermination and prevention in the Results section (see also Discussion 4.2).

Fig. 4 shows the entire causal account of our example transition. Intuitively, in this simple example, $x_{t-1} = \{ \text{OR} = 1 \}$ has the actual effect $y_t = \{ \text{OR} = 1 \}$ and is also the actual cause of $y_t = \{ \text{OR} = 1 \}$, and the same for $x_{t-1} = \{ \text{AND} = 0 \}$ and $y_t = \{ \text{AND} = 0 \}$. Nevertheless, there is also a causal link $x_{t-1} \leftarrow y_t$, between the 2nd-order occurrence $y_t = \{ \text{OR}, \text{AND} = 10 \}$ and its actual cause $x_{t-1} = \{ \text{OR}, \text{AND} = 10 \}$, which is irreducible to its parts, as shown in Fig. 3D (right). The complementary link, however, does not exist, $x_{t-1} \rightarrow y_t$, it is reducible (Fig. 3C, right). The causal account $C$ shown in Fig. 4 provides a complete causal explanation for “what happened” and “what caused what” in the transition $x_{t-1} \prec y_t$.

**2.2 Irreducibility of the causal account**

The previous section defined the causal account of a particular transition $x_{t-1} \prec y_t$ with $X_{t-1}, Y_t \subseteq S$. A causal account is the set of all causal links within a given transition. Occurrences without causes or effects, or links that are reducible to their parts are not included in the causal account. As with potential causation (Oizumi et al., 2014; Albantakis and Tononi, 2015), the principle of integration can be applied not only to individual causal links, but also to the causal account as a whole. That is, given a transition, is it best characterized by a single, integrated causal account, or by multiple, independent causal accounts? In other words, to what extent are the observations $x_{t-1}, y_t$ and the corresponding transition $x_{t-1} \prec y_t$ irreducible to their parts?

To test the irreducibility of observation $x_{t-1}$ and its set of actual effects in $y_t$, the transition $x_{t-1} \prec y_t$ is partitioned into independent parts in the same way that $x_{t-1} \rightarrow y_t$ is partitioned when assessing $\alpha_e(x_{t-1}, y_t)$, resulting in a set of possible partitions $P(x_{t-1} \rightarrow y_t)$. We then define the irreducibility of $x_{t-1}$ as the total strength of actual effects (causal links of the form $x_{t-1} \rightarrow y_t$) lost from the causal account $C$ due to the MIP:

$$\Delta_e(x_{t-1} \prec y_t) = \sum_{x-y \in C} \alpha_e^{\max}(x) - \sum_{x-y \in MIP} \alpha_e^{\max}(x)_{MIP}$$

(10)

In the same way, the irreducibility of observation $y_t$ and its set of actual causes in $x_{t-1}$ is defined as the total strength of actual causes (causal links of the form $x_{t-1} \leftarrow y_t$) lost from the causal account $C$ due to the MIP:

$$\Delta_c(x_{t-1} \prec y_t) = \sum_{x-y \in C} \alpha_c^{\max}(y) - \sum_{x-y \in MIP} \alpha_c^{\max}(y)_{MIP}$$

(11)

where the MIP is again the partition that makes the least difference out of all possible partitions $P(x_{t-1} \leftarrow y_t)$.

As for causal links, the irreducibility of a single element observation $x_{t-1}$ or $y_t$ reduces to $\alpha_e^{\max}$ of its one actual effect $y_t$, or $\alpha_c^{\max}$ of its one actual cause $x_{t-1}$, respectively.

By considering the union of possible partitions, $P(x_{t-1} \prec y_t) = P(x_{t-1} \rightarrow y_t) \cup P(x_{t-1} \leftarrow y_t)$, we can moreover assess the overall irreducibility of the transition $x_{t-1} \prec y_t$. A transition $x_{t-1} \prec y_t$ is reducible if there is a partition $P \in P(x_{t-1} \prec y_t)$ such that the total strength of causal links in $C(x_{t-1} \prec y_t)$ is unaffected by the partition. Based on this notion we define the irreducibility of a transition $x_{t-1} \prec y_t$ as:

$$\Delta_e(x_{t-1} \prec y_t) = \sum \alpha_e^{\max}(C) - \sum \alpha_e^{\max}(C_{MIP})$$

(12)

where

$$\sum \alpha_e^{\max}(C) = \sum_{x-y \in C} \alpha_e^{\max}(x) + \sum_{x-y \in C} \alpha_e^{\max}(y)$$

is a summation over the strength of all causal links in the causal account $C(x_{t-1} \prec y_t)$, and the same for the partitioned causal account $C_{MIP}$. 

8
Fig. 4 shows the causal account \( C_{\text{MIP}} \) of the transition \( x_{t-1} \prec y_t \) with \( x_{t-1} = y_t = \{ \text{OR, AND} = 10 \} \) under its MIP into \( m = 2 \) parts with \( x_{1,t-1} = \{ \text{OR} = 1 \}, x_{2,t-1} = \{ \text{AND} = 1 \} \), \( y_{t1} = \{ \text{OR} = 1 \} \), and \( y_{t2} = \{ \text{AND} = 0 \} \), with \( A(x_{t-1} < y_t) = 0.17 \). This is the causal strength that would be lost if we treated \( x_{t-1} < y_t \) as two separate transitions instead of a single one.

The measures defined in this section provide the means to exhaustively assess “what caused what” in a system transition \( s_{t-1} \prec s_t \), to evaluate more parsimonious causal explanations against the complete causal account of \( s_{t-1} \prec s_t \), and to evaluate the strength of specific causal links of interest in a particular context (background conditions).

\[
\begin{align*}
\text{Causal account } C(x_{t-1} < y_t) & \\
 x_{t-1} = \{ \text{OR, AND} = 10 \} \prec y_t = \{ \text{OR, AND} = 10 \}
\end{align*}
\]

\[
\begin{align*}
& x \rightarrow y^* \\
& a_c^{\text{max}} = 0.415 \text{ bits} \\
& \text{actual effect} \\
& \sum a_c^{\text{max}}(C) = 1.83 \text{ bits} \\
& \mathcal{A}_c = 0 \text{ bits} \\
& \mathcal{A} = \mathcal{A}_c = 0.17 \text{ bits}
\end{align*}
\]

\[
\begin{align*}
& x^* \leftarrow y \\
& a_c^{\text{max}} = 0.415 \text{ bits} \\
& \text{actual cause} \\
& \sum a_c^{\text{max}}(C) = 1.83 \text{ bits} \\
& \mathcal{A}_c = 0 \text{ bits} \\
& \mathcal{A} = \mathcal{A}_c = 0.17 \text{ bits}
\end{align*}
\]

**Fig. 4: Irreducible Causal Account.** There are two 1\(^{st}\)-order occurrences with actual effects and actual causes. In addition, the 2\(^{nd}\)-order occurrence \( y_t = \{ \text{OR, AND} = 10 \} \) has an actual cause \( x_{t-1} = \{ \text{OR, AND} = 10 \} \). A partition of the transition along the MIP destroys this 2\(^{nd}\)-order causal link, leading to \( \mathcal{A} = 0.17 \) bits.

3. Results

In the following, we will present a series of examples to illustrate the quantities and objects defined in the theory section and address several dilemmas taken from the literature on actual causation.

3.1 The same transition can have different causal accounts: disjunction, conjunction, biconditional, and prevention

Fig. 5 shows 4 different types of logic gates with two inputs each, all transitioning from the input state \( x_{t-1} = \{ \text{AB} = 11 \} \) to the output state \( y_t = \{ \text{C} = 1 \}, \{ \text{D} = 1 \}, \{ \text{E} = 1 \} \) or \( \{ \text{F} = 1 \} \). From a dynamical point of view, without taking the causal structure of the mechanisms into account, the same occurrences are happening in all four situations. However, analyzing the causal accounts of these transitions reveals differences in the number and type of irreducible occurrences and their actual causes and effects.

The first example (Fig. 5A – OR gate), is a case of symmetric overdetermination (Pearl, 2000, Chapter 10); each input to C would have been sufficient for \( \{ \text{C} = 1 \} \), yet both \( \{ \text{A} = 1 \} \) and \( \{ \text{B} = 1 \} \) occurred at \( t - 1 \). By contrast, in the second example (Fig. 5B – AND gate), both \( \{ \text{A} = 1 \} \) and \( \{ \text{B} = 1 \} \) are necessary for \( \{ \text{D} = 1 \} \). These two examples (Fig. 5A and B) are often referred to as the disjunctive and conjunctive versions of the ‘forest-fire’ example (Halpern and Pearl, 2005; Halpern, 2015, 2016), where lightning and/or a match being dropped result in a forest fire. The main question is whether, assuming both the lightning and the match being dropped occurred, the actual cause of the forest fire was a) one of the two, b) either, or c) both together. Here we argue that separating actual effects from actual causes and acknowledging compositional causes and effects can resolve these situations.

In the disjunctive case (Fig. 5A), each of the inputs to C has an actual effect, \( \{ \text{A} = 1 \} \rightarrow \{ \text{C} = 1 \} \) and \( \{ \text{B} = 1 \} \rightarrow \{ \text{C} = 1 \} \), as they raise the probability of \( \{ \text{C} = 1 \} \) compared to its unconstrained probability. Yet, by the causal exclusion principle, the occurrence \( \{ \text{C} = 1 \} \) can only have one actual cause. The link between \( \{ \text{C} = 1 \} \) and \( \{ \text{AB} = 11 \} \) is reducible with \( \alpha_c = 0 \), so \( \{ \text{AB} = 11 \} \) is not the actual cause of \( \{ \text{C} = 1 \} \). Instead, either \( \{ \text{A} = 1 \} \), or \( \{ \text{B} = 1 \} \) is the actual cause, but not both. Note that which of the two inputs it remains undetermined, since they are perfectly symmetric in this example (compare to Results 3.3 below).

In the conjunctive example (Fig. 5B), each input alone has an actual effect, \( \{ \text{A} = 1 \} \rightarrow \{ \text{C} = 1 \} \) and \( \{ \text{B} = 1 \} \rightarrow \{ \text{C} = 1 \} \) (with higher strength than in the disjunctive case), but here also the 2\(^{nd}\)-order occurrence of both inputs together has an actual effect, \( \{ \text{AB} = 11 \} \rightarrow \{ \text{D} = 1 \} \). Thus, there is a composition of actual effects. Again, the occurrence \( \{ \text{D} = 1 \} \) can only have one actual cause; here the 2\(^{nd}\)-order cause \( \{ \text{AB} = 11 \} \) of \( \{ \text{D} = 1 \} \), with \( \alpha_c^{\text{max}} = 2.0 \) (note that \( \{ \text{A} = 1 \} \) and \( \{ \text{B} = 1 \} \) are not separate causes, as was concluded instead by (Halpern, 2015)).

The importance of high-order occurrences is emphasized by the third example (Fig. 5C), where E is a “logical biconditional” (an XNOR) of its two inputs. In this case, the individual occurrences \( \{ \text{A} = 1 \} \) and \( \{ \text{B} = 1 \} \) by themselves make no difference in bringing about \( \{ \text{E} = 1 \} \); their effect ratios are zero. For this reason, they cannot have actual effects and cannot be actual causes. Only the 2\(^{nd}\)-order occurrence \( \{ \text{AB} = 11 \} \) specifies \( \{ \text{E} = 1 \} \), which is its actual effect, \( \{ \text{AB} = 11 \} \rightarrow \{ \text{E} = 1 \} \). Likewise, \( \{ \text{E} = 1 \} \) only specifies the 2\(^{nd}\)-order occurrence \( \{ \text{AB} = 11 \} \), not its parts taken separately, thus \( \{ \text{AB} = 11 \} \rightarrow \{ \text{E} = 1 \} \).
Fig. 5: Four dynamically identical transitions can have different causal accounts. Shown are the transitions (top) and their respective causal accounts (bottom). The causal links lost through the MIP are indicated in red.

Note that the causal strength in this example is lower than in the case of the AND-gate, since \( \{D = 1\} \) is less likely to begin with than \( \{E = 1\} \).

The last example (Fig. 5D) demonstrates a case in which the transition \( x_{t-1} \leftarrow y_t \) is reducible \((\mathcal{A}(x_{t-1} < y_t) = 0\)\). In this case, \( \{B = 1\} \rightarrow \{F = 1\} \) and \( \{B = 1\} \leftarrow \{F = 1\} \); however, \( \{A = 1\} \) does not have an actual effect and is not an actual cause. This example can be seen as a case of prevention: \( \{B = 1\} \) causes \( \{F = 1\} \), which prevents any effect of \( \{A = 1\} \). In a popular narrative accompanying this example, \( \{A = 1\} \) is an assassin putting poison in the King’s tea, while a bodyguard administers an antidote \( \{B = 1\} \), preventing the King’s death \( \{F = 1\} \) (Halpern, 2016)\(^6\). Since \( \{A = 1\} \) does not contribute to any causal links, partitioning element A away from B and F, which is equivalent to replacing element A by ‘noise’ (maximum entropy), does not make a difference to the causal account and thus \( \mathcal{A}(x_{t-1} < y_t) = 0 \). Note that the causal account is state dependent: for a different transition, element A may have an actual effect or contribute to an actual cause, and the causal account may become irreducible.

Taken together, the above examples demonstrate that the causal account and the causal strength of individual causal links within the account capture differences in sufficiency and necessity of the various occurrences in their respective transitions. In particular, not all occurrences at \( t - 1 \) with actual effects end up being actual causes of occurrences at \( t \).

\(^6\) Note however that this causal model is equivalent to an OR-gate, as can be seen by switching the state labels of A from ‘0’ to ‘1’ and vice versa. The discussed transition would correspond to the case of one input to the OR-gate being ‘1’ and the other ‘0’. Since the OR-gate switches on (‘1’) in this case, the ‘0’ input has no effect and is not a cause.

### 3.2 Linear threshold units

A generalization of simple, linear logic gates, such as OR- and AND-gates, are binary linear threshold units (LTUs). Given \( n \) equivalent inputs \( x_{t-1} \), a single LTU \( Y_t \) will turn on (‘1’) if the number of inputs in state ‘1’ exceeds a given threshold \( k \),

\[
p(Y_t = 1|x_{t-1}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} x_{i,t-1} \geq k, \\ 0 & \text{if } \sum_{i=1}^{n} x_{i,t-1} < k. \end{cases}
\]

LTUs are of great interest, for example, in the field of neural networks, since they comprise one of the simplest model mechanisms for neurons, capturing the notion that a neuron fires if it received sufficient synaptic inputs. One example is a Majority-gate, which outputs ‘1’ if more than half of its inputs are ‘1’.

Fig. 6 displays the causal account of a Majority-gate \( M \) with 4 inputs for the transition \( x_{t-1} = \{ABC = 1110\} \rightarrow y_t = \{M = 1\} \). All of the inputs in state ‘1’, as well as their high-order occurrences, have actual effects on \( \{M = 1\} \). Occurrence \( \{D = 0\} \), however, does not work towards bringing about \( \{M = 1\} \); it reduces the probability for \( \{M = 1\} \) and thus does not contribute to any actual effects or the actual cause. As with the AND-gate in the previous section, there is composition of actual effects in the causal account. Yet, there is only one actual cause, \( \{ABC = 111\} \rightarrow \{M = 1\} \). In this case, it happens to be that the 3rd-order occurrence \( \{ABC = 111\} \) is minimally sufficient for \( \{M = 1\} \)—no smaller set of input elements would suffice. Note however, that the actual cause \( x_{t-1}^M (\{M = 1\}) \) is not determined based on sufficiency, but because \( \{ABC = 111\} \) is the set of elements maximally constrained by the occurrence \( \{M = 1\} \). Nevertheless, causal analysis as illustrated here will always identify a minimally sufficient set of input elements as the actual cause of an LTU, for any
number of inputs $n$ and any threshold $k$. Furthermore, any occurrence of all ‘1’ elements with size less than or equal to the threshold $k$ will be irreducible, with the LTU as their actual effect.

$$x_{t-1} = (ABCD = 1110) \prec y_t = \{M = 1\}$$

$$\mathcal{A} = \mathcal{A}_e = 0 \text{ bits}$$

**Theorem 1:** For a transition $x_{t-1} \prec y_t$, where $y_t$ is a single LTU with $n$ inputs and a threshold of $k$, and $x_{t-1}$ is the set of inputs to the LTU, the following holds:

1. The actual cause of $\{Y_t = 1\}$ is $\{X_{t-1} = x_{t-1}\}$, such that $|x_{t-1}| = k$ and min($x_{t-1}$) = 1.
2. If min($x_{t-1}$) = 1 and $|x_{t-1}| \leq k$; then $\{Y_t = 1\}$ is the actual effect of $\{X_{t-1} = x_{t-1}\}$; otherwise $x_{t-1}$ has no actual effect, it is reducible.

Proof: See Supplementary Proof 1.

Note that an LTU in the off (‘0’) state, $\{Y_t = 0\}$, has equivalent results with the role of ‘0’ and ‘1’ reversed, and a threshold of $n - k$. In the case of overdetermination, e.g. the transition $x_{t-1} = (ABCD = 1111) \prec y_t = \{M = 1\}$, where all inputs to the Majority-gate are ‘1’, the actual cause will again be a subset of 3 input elements in state ‘1’.

However, which of the possible sets remains undetermined due to symmetry, just as in the case of the OR-gate in Fig. 5A.

The transition $x_{t-1} = \{ABCD = 1110\} \prec y_t = \{M = 1\}$ shown in Fig. 6 is reducible with $\mathcal{A}(x_{t-1} \prec y_t) = 0$, since input element D does not contribute to the causal account. Consider instead the causal account of the transition $x_{t-1} = \{ABC = 111\} \prec y_t = \{M = 1\}$, where $\{D = 0\}$ at $t - 1$ is treated as a background condition (Supplementary Fig. S1). Doing so results in a causal account with the same causal links but higher causal strengths. This captures the intuition that A, B, and C’s ‘yes votes’ are more important if it is already determined that D will ‘vote no’.

### 3.3 Disjunction of conjunctions

Another case often considered in the actual causation literature is a disjunction of conjunctions, that is, an OR-operation over two or more AND-operations. In the general case, a disjunction of conjunctions is an element $Y_t$ that is a disjunction of $k$ conditions, each of which is a conjunction of $n_j$ input elements $x_{i,j,t-1}$, for $i = 1 \ldots n_j$, and $j = 1 \ldots k$,

$$p(Y_t = 1|\mathcal{X}_{t-1}) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n_j} x_{i,j,t-1} < n_j \forall j \\ 1 & \text{otherwise} \end{cases}$$

Here we consider a simple example, $(A \land B) \lor C$ (Fig. 7). The debate over this example is mostly concerned with the type of transition shown in Fig. 7: $x_{t-1} = \{ABC = 101\} \prec y_t = \{D = 1\}$, and the question whether $\{A = 1\}$ is a cause of $\{D = 1\}$ even if $B = 0$. Again, the distinction between actual effects and actual causes, as well as the quantitative assessment of actual causes and actual effects, can help to resolve issues of actual causation in this type of examples.

As shown in Fig. 7, with respect to actual effects, both $\{A = 1\} \rightarrow \{D = 1\}$ and $\{C = 1\} \rightarrow \{D = 1\}$, with $\{C = 1\}$ having a stronger actual effect. With respect to actual causes, only $\{C = 1\} \leftarrow \{D = 1\}$, being the maximally irreducible cause with $\mathcal{A}^\text{max}(\{D = 1\}) = 0.678$.

When judging the actual effect of $\{A = 1\}$ at $t - 1$ within the transition $x_{t-1} = \{ABC = 101\} \prec y_t = \{D = 1\}$, B is assumed to be undetermined. By itself, the occurrence $\{A = 1\}$ does raise the probability of occurrence $\{D = 1\}$, and thus $\{A = 1\} \rightarrow \{D = 1\}$. Note however, that this transition is reducible with $\mathcal{A}(x_{t-1} \prec y_t) = 0$, since B does not contribute to the causal account. The occurrence $\{C = 1\}$ prevents $\{B = 0\}$ from having an actual effect (if C were 0, D would be 0 and $\{B = 0\}$ would have an actual effect).

---

9 Halpern (2016) considers a voting scenario with eleven voters for two candidates, which is equivalent to a Majority-gate with 11 inputs. In the case of all 11 inputs in state ‘1’, he arrives at the conclusion that any subset of 6—any minimally sufficient set—is a cause for $y_t = \{M = 1\}$ (or ‘Suzy winning’ in his example), according to the modified HP definition of an actual cause. Note, however, that the modified HP account does not generally identify minimally sufficient sets as the actual cause. Given 10 voters, all voting for ‘1’, the modified HP account would identify any subset of 5 voters as the actual cause (the number of ‘1s’ that would have to be switched to change the outcome of M), while still 6 votes are necessary and minimally sufficient for a majority ($M = 1$).

10 One story accompanying this example is that “a prisoner dies either if A loads B’s gun and B shoots, or if C loads and shoots his gun... A loads B’s gun, B does not shoot, but C does load and shoot his gun, so that the prisoner dies” (Hopkins and Pearl, 2003; Halpern, 2016).
effect). However, if \( \{B = 0\} \) is taken as a fixed background condition, the transition \( x_{t-1} = \{AC = 11\} < y_t = \{D = 1\} \) is also reducible, since with \( \{B = 0\} \) fixed, \( \{A = 1\} \) does not have an actual effect anymore (Fig. 2A). In this case, the background condition \( \{B = 0\} \) prevents \( \{A = 1\} \) from having any effect. The only irreducible transition with \( \mathcal{A} = 2.0 \) is \( x_{t-1} = \{C = 1\} < y_t = \{D = 1\} \) with \( \{AB = 10\} \) as a fixed background condition, which may thus be interpreted as the most parsimonious explanation for what happened.

![Fig. 7: Disjunction of two conjunctions (A \& B) \lor C.](image)

The results from this example extend to the general case of disjunctions of conjunctions. In the situation where \( y_t = 1 \), the actual cause of \( y_t \) is the minimum set of sufficient elements. If multiple conjunctive conditions are satisfied, the actual cause of \( y_t \) is thus the smallest such condition (see Supplementary Fig S2B for such an example of asymmetric overdetermination). At \( t - 1 \), any 1\textsuperscript{st}-order occurrence in state ‘1’, as well as any high-order occurrence of such elements that does not overdetermine \( y_t \), has an actual effect. This includes any occurrence in state all ‘1’ that contains only elements from exactly one conjunction, as well as any high-order occurrence of elements across conjunctions, which do not fully contain any specific conjunction.

If instead \( y_t = 0 \), then the actual cause of \( y_t \) is an occurrence that contains a single element in state ‘0’ from each conjunctive condition. At \( t - 1 \), any occurrence in state all ‘0’ that does not overdetermine \( y_t \) has an actual effect, which is any all ‘0’ occurrence that does not contain more than one element from any conjunction.

These results are formalized by the following theorem.

**Theorem 2.** For a transition \( x_{t-1} < y_t \), where \( y_t \) is a single element observation that is a disjunction of \( k \) conditions, each of which is a conjunction over \( n_j \) elements, and \( X_{t-1} = \{f_{j,t-1}\} = \{x_{j,t-1}\} \) is the set of all \( n = \sum_{j=1}^{k} n_j \) inputs. Considering occurrences \( x_{t-1} \) with \( c_j \) elements from each conjunction, the following holds:

1. If \( y_t = 1 \):
   a. The actual cause of \( \{Y_t = 1\} \) is \( \{X_{t-1} = f_{j,t-1}\} \), such that \( \min(x_{j,t-1}) = 1 \) and \( \min(x_{j,t-1}) = 1 \Rightarrow n_j \leq n_i \).
   b. If \( \min(x_{t-1}) = 1 \) and \( c_j = n_j \Rightarrow c_j = 0, \forall j, i \neq j \), then \( \{Y_t = 1\} \) is the actual effect of \( \{X_{t-1} = x_{t-1}\} \), otherwise \( x_{t-1} \) has no actual effect, it is reducible.

2. If \( y_t = 0 \):
   a. The actual cause of \( \{Y_t = 0\} \) is \( \{X_{t-1} = x_{t-1}\} \), such that \( \max(x_{t-1}) = 0 \) and \( c_j \leq 1 \).
   b. If \( \max(x_{t-1}) = 0 \) and \( c_j \leq 1 \), then \( \{Y_t = 0\} \) is the actual effect of \( \{X_{t-1} = x_{t-1}\} \), otherwise \( x_{t-1} \) is reducible and has no actual effect.

Proof: See Supplementary Proof 2.

### 3.4 Complicated voting

As already demonstrated in the examples in Fig. 5C and D, the proposed causal analysis is not restricted to linear update functions or combinations thereof. Fig. 8 depicts an example transition featuring a complicated, nonlinear update function. This specific example is taken from (Halpern, 2015, 2016): If A and B agree, F takes their value, if B, C, D, and E agree, F takes A’s value, otherwise majority decides. The transition of interest is \( x_{t-1} = (ABCDE = 11000) < y_t = \{F = 1\} \).

According to Halpern (Halpern, 2015), intuition suggests that \( \{AB = 11\} \sim \{F = 1\} \). \( \{AB = 11\} \) is also the minimally sufficient occurrence in the transition that determines \( \{F = 1\} \). Indeed, this is the result of the present causal analysis of the transition (Fig. 8). In addition, \( \{A = 1\} \sim \{F = 1\} \), \( \{B = 1\} \sim \{F = 1\} \), \( \{AB = 11\} \sim \{F = 1\} \), and \( \{ACDE = 1000\} \sim \{F = 1\} \) all contribute to the causal account. Both \( \{AB = 11\} \) and \( \{ACDE = 1000\} \) completely determine that \( \{F = 1\} \) will occur. Nevertheless, according to the exclusion principle, \( \{AB = 11\} \) is the actual cause of \( \{F = 1\} \) rather than \( \{ACDE = 1000\} \), as it is the minimal occurrence with \( \alpha_c(x_{t-1}; y_t) = \alpha_{max}(y_t) = 1.0 \).

![Fig. 8: Complicated voting.](image)

### 3.5 Noise and probabilistic elements

The examples so far involved deterministic update functions. Probabilistic accounts of causation are closely related to counterfactual accounts (Paul and Hall, 2013). Nevertheless, certain problem cases only arise in probabilistic settings (e.g. Fig. 9B). The present causal analysis can be applied equally to probabilistic as well as
deterministic systems, as long as the system’s transition probabilities satisfy conditional independence (Eqn. 1). No separate, probabilistic calculus for actual causation is required.

In the simplest case, where noise is added to a deterministic transition \( x_{t-1} < y_t \), the noise will generally decrease the strength of the causal links in the transition. Fig. 9 shows the causal account of the transition \( x_{t-1} = \{ A = 1 \} < y_t = \{ N = 0 \} \), where \( N \) is the slightly noisy version of a COPY-gate. In this example, both \( \{ A = 1 \} \rightarrow \{ N = 1 \} \) and \( \{ A = 1 \} \leftarrow \{ N = 1 \} \). The only difference with the equivalent deterministic case is that the causal strength \( a_{\text{max}} = 0.848 \) is lower than in the deterministic case where \( a_{\text{max}} = 1.06 \).

Note that in this probabilistic setting, the actual cause \( \{ A = 1 \} \) by itself is not sufficient to determine \( \{ N = 0 \} \). Nevertheless, \( \{ A = 1 \} \) makes a positive difference in bringing about \( \{ N = 1 \} \), and this difference is irreducible, so the causal link exists within the transition.

The transition \( x_{t-1} = \{ A = 1 \} < y_t = \{ N = 0 \} \) has no counterpart in the deterministic case (where it would be impossible by violating the existence principle). The result of the causal analysis is that there are no irreducible causal links within the given transition, which is therefore reducible with \( A(x_{t-1} < y_t) = 0 \). One interpretation is that the actual cause of \( \{ N = 0 \} \) must lie outside of the system, such as a missing latent variable. Another interpretation is that the actual cause for \( \{ N = 0 \} \) is genuine ‘physical noise’. In any case, the proposed account of actual causation is sufficiently general to cover both deterministic and as well as probabilistic systems.

![Diagram](image)

**Fig. 9: Probabilistic elements.** While the transition shown in (A) does have a deterministic equivalent, the transition shown in (B) would be impossible in the deterministic case.

### 3.6 Simple classifier

As a final example, we consider a transition with a multi-element observation \( y_t \)—in particular, 3 elements A, B, and C providing input to 3 different classifier elements D, S, and L. Element D is a ‘dot-detector’; it outputs ‘1’ if exactly one of the 3 inputs is in state ‘1’. Element S is a ‘segment-detector’; it outputs ‘1’ if input states \( \{ ABC = 110 \} \) and \( \{ ABC = 011 \} \). Element L detects lines, that is, \( \{ ABC = 111 \} \).

Fig. 10 shows the causal account of the specific transition \( x_{t-1} = \{ ABC = 001 \} < y_t = [DSL = 100] \). In this case, only a few occurrences \( x_{t-1} \) have actual effects, but all possible occurrences \( y_t \) are irreducible with their own actual cause. The occurrence \( \{ C = 1 \} \) by itself, for example, has no actual effect. This may be initially surprising since D is a dot detector and \( \{ C = 1 \} \) is supposedly a dot. However, the specific configuration of the entire input set is necessary to determine \( \{ D = 1 \} \) (a dot is only a dot if the other inputs are ‘0’). Consequently, \( \{ ABC = 001 \} \rightarrow \{ D = 1 \} \) and also \( \{ ABC = 001 \} \leftarrow \{ D = 1 \} \). By contrast, the occurrence \( \{ A = 0 \} \) is sufficient to determine \( \{ L = 0 \} \) and raises the probability of \( \{ D = 1 \} \); the occurrence \( \{ B = 0 \} \) is sufficient to determine \( \{ S = 0 \} \) and \( \{ F = 0 \} \) and also raises the probability of \( \{ D = 1 \} \). We thus get the following causal links from single element occurrences in the system: \( \{ A = 0 \} \rightarrow \{ DL = 10 \} \), \( \{ A = 0 \} \leftarrow \{ L = 0 \} \), \( \{ B = 0 \} \rightarrow \{ DSL = 100 \} \) and \( \{ B = 0 \} \leftarrow \{ S = 0 \} \) (note that \( \{ B = 0 \} \) would be an equivalent actual cause of \( \{ L = 0 \} \) with the same \( a_{\text{max}} = 1.415 \).

In addition, all high-order occurrences \( y_t \) are irreducible, each having their own actual cause above those of their parts. The actual cause identified for these high-order occurrences can be interpreted as the ‘strongest’ shared cause of the elements in the occurrence, for example \( \{ B = 0 \} \leftarrow \{ DS = 10 \} \). While only the occurrence \( \{ ABC = 001 \} \) is sufficient to determine \( \{ DS = 10 \} \), this candidate causal link is reducible, because \( \{ DS = 10 \} \) does not constrain the state of ABC any more than \( \{ D = 1 \} \) by itself. In fact, the occurrence \( \{ S = 0 \} \) does not constrain the state of AC at all. Thus \( \{ ABC = 001 \} \) and all other candidate causes of \{DS = 10\} that include these elements are either reducible (because their causal link can be partitioned with \( a_{\text{max}} = 0 \)) or excluded (because there is a smaller set of elements whose causal strength is at least as high). In this example, \( \{ B = 0 \} \) is the only irreducible shared cause of \( \{ D = 1 \} \) and \( \{ S = 0 \} \), and thus also the actual cause of \( \{ DS = 10 \} \).

![Diagram](image)

**Fig. 10: Simple classifier.** Element D is a ‘dot-detector’, S a ‘segment-detector’, and L a ‘line-detector’ (see text).
4. Discussion

In this article, we presented a principled, comprehensive formalism to assess actual causation given a transition between two subsequent observations in a discrete dynamical system. The proposed framework provides a complete causal account of all causal links between occurrences at \( t - 1 \) and \( t \), based on five principles – existence, composition, information, integration, and exclusion. In what follows, we review specific features and limitations of our approach, discuss how the results relate to intuitive notions about actual causation and causal explanation, and highlight some of the main differences with previous proposals aimed at operationalizing the notion of actual causation. Specifically, our framework considers all possible counterfactual states and makes it possible to assess the actual strength of causal links. Second, it distinguishes between actual causes and effects, which are considered separately. Third, it allows for causal composition, in the sense that first- and high-order occurrences can have their own causes and effects within the same transition, as long as they are irreducible. As demonstrated in the results section, the proposed formalism is generally applicable to a vast range of mechanistic relations. By enforcing the assessment of actual causation in fully specified model systems, this framework also avoids several paradoxes that can arise from narrative examples loosely tied to a causal model.

4.1 Testing all counterfactuals with equal probability

From a single transition \( x_{t-1} < y_t \), i.e. an observed sequence of two states, it is neither possible to infer causal relations between the elements involved, nor to identify the mechanisms underlying such causal interactions (see Fig. 5). Moreover, observing the same transition over and over again does not automatically reveal a causal link between \( x_{t-1} \) and \( y_t \). In order to infer a causal dependency between \( x_{t-1} \) and \( y_t \), one has to assess alternative states (counterfactuals) of both sets of elements \( X_{t-1} \) and \( Y_t \).

The structural modeling framework of system interventions introduced by Pearl (Pearl, 2000) provides a basis for formalizing the notion of counterfactual inference. In the simplest case, counterfactual approaches to actual causation are based on the “but-for” test (Halpern, 2016): \( c \) is a cause of \( e \) if \( \neg c \implies \neg e \) (“but for \( c \), \( e \) would not have happened”). In larger causal models in which this condition is typically context dependent (on variables \( Z \)), more sophisticated counterfactual approaches differ depending on the various contingencies \( Z = z \) under which the “but-for” test is applied. In general, if one such counterfactual state \( \neg \exists c, z \) exists, \( c \) is identified as a cause of \( e \) (Pearl, 2000; Halpern, 2016).

By contrast, the present framework requires that interventions set a discrete dynamical system \( S \) into all its possible states, including those that may not spontaneously occur during the system’s dynamical evolution. This is because a discrete dynamical system \( S \) and the mechanisms of its elements are only specified fully by the transition probabilities of all possible system states (Eqn. 1) (Ay and Polani, 2008; Oizumi et al., 2014). For example, element \( E \) in Fig. 5C, an XNOR, can only be distinguished from the AND-gate \( D \) in Fig. 5B, and \( F \) in Fig. 5D, if all its transition probabilities (input-output relations) from every possible state to the next state are given. Once the underlying mechanisms are specified based on all possible transition probabilities, causal relations can be quantified in probabilistic terms (Ay and Polani, 2008; Oizumi et al., 2014) even within a single transition \( x_{t-1} < y_t \), i.e. in the context of actual causation (Glennan, 2011; Korb et al., 2011). In this way, the present approach to actual causation is applicable to both deterministic and probabilistic systems and explicitly captures the notion of causal specificity within the scope of the transition of interest (Woodward, 2010; Hoel et al., 2013; Oizumi et al., 2014).

Accordingly, actual causes and effects of occurrences within a transition \( x_{t-1} < y_t \), which reflect mechanistic dependencies between system variables, are identified based on interventional conditional probability distributions derived from the set of all possible transition probabilities (Eqn. 1) (Ay and Polani, 2008; Balduzzi and Tononi, 2008; Oizumi et al., 2014). To correctly capture the strength of the causal constraints \( x_{t-1} \) exerts over \( y_t \) and vice versa, unconfounded by constraints due to other endogenous or external variables and previous time steps, we explicitly require that the distribution over states of \( X_{t-1} \) be maximum entropy. In other words, the probability of an observation \( x_{t-1} \) should have no bearing on the question whether \( x_{t-1} \) had an effect on or was a cause of \( y_t \), once \( x_{t-1} \) actually happened\(^{11}\). Related measures of information flow in causal networks (Ay and Polani, 2008) and causal information (Korb et al., 2011), by contrast, remain agnostic about the assumed distribution over states of \( X_{t-1} \).

4.2 Distinguishing actual effects and actual causes

An implicit assumption commonly made about (actual) causation is that the relation between cause and effect is bidirectional: if occurrence \( C = c \) had an effect on occurrence \( E = e \), then \( c \) is assumed to be a cause of \( e \). Thus, most approaches aimed at identifying the actual cause of an occurrence \( e \) ask whether a manipulation of the candidate cause in a given context changes the state of \( e \), or

\(^{11}\) Likewise, which other occurrences in \((x/x)_{t-1}\) happened and how likely they were to happen should not matter for determining the strength with which \( x_{t-1} \) constrains \( y_t \). This is why we use the product distributions (Eqn. 2 and 3), assuming independent, maximum entropy inputs to each element in \( Y_t \).
its probability (e.g., Woodward, 2004; Korb et al., 2011; Halpern, 2016). In probabilistic accounts this view is manifested in the notion that causes should raise the probability of their effects, such as the requirement that \( p(e|do(C = c)) > p(e) \) (Good, 1961; Suppes, 1970; Eells and Sober, 1983; Pearl, 2009)\(^{12}\). In the present framework, \( p(e|do(C = c)) > p(e) \) is captured by the effect ratio \( \rho_e(x_{t-1}, y_t) \) (Eqn. 4), which also takes into account all possible counterfactuals (all \( x \in \Omega_x \)). However, even in deterministic systems, the fact that an occurrence \( e \), by itself, raises the probability of an occurrence \( e \), does not necessarily determine that \( E = e \) will actually occur (Paul and Hall, 2013). Consider, for example, the transition shown in Fig. 7: by itself, the occurrence \( \{A = 1\} \) raises the probability of \( \{D = 1\} \) (\( \rho_e(x_{t-1}, y_t) = a_e(x_{t-1}, y_t) > 0 \)). Nevertheless, in the example transition, \( \{A = 1\} \) is neither necessary nor sufficient for \( \{D = 1\} \): if not for \( \{C = 1\} \), \( D \) would be in state ‘0’ instead of ‘1’. Accordingly, it has been argued that \( \{A = 1\} \) should not be considered a cause of \( \{D = 1\} \) (Halpern, 2015). Our approach clarifies the causal account by explicitly distinguishing effects from causes and considering both separately. Actual effects are identified from the perspective of occurrences at \( t - 1 \), whereas actual causes are identified from the perspective of occurrences at \( t \). In Fig. 7, while both \( \{A = 1\} \) and \( \{C = 1\} \) have an actual effect on \( \{D = 1\} \), \( \{C = 1\} \) is the maximally irreducible and thus actual cause of \( \{D = 1\} \) (following the exclusion principle). In general, the existence of an actual effect \( x_{t-1} \rightarrow y_t \) does not imply that the complementary actual cause \( x_{t-1} \leftarrow y_t \) also exists and vice versa.

Here we propose that a comprehensive causal understanding of a given transition is provided by its complete causal account \( \mathcal{C} \) (Definition 2.3), including both actual effects and actual causes, i.e. all maximally irreducible causal links of direction \( x_{t-1} \rightarrow y_t \) and \( x_{t-1} \leftarrow y_t \). In this way, the causal account provides a comprehensive explanation of “what happened” in terms of “what caused what”, which acknowledges causal exclusion and still provides all causally relevant details. As an example, Fig. 5 illustrates how the causal account describes and distinguishes four dynamically identical transitions. Including both actual causes and effects into the causal account can moreover resolve ambiguities in cases of redundant causation, such as overdetermination, prevention, and prohibition (Hitchcock, 2007; Paul and Hall, 2013). As an example, consider again the system shown in Fig. 7, but now transitioning from \( x_{t-1} = \{ABC = 111\} \) to \( y_t = \{D = 1\} \) (Fig. 8B). In this case of asymmetric overdetermination, from the perspective of \( t - 1 \), the occurrences \( \{A = 1\}, \{B = 1\}, \{AB = 11\}, \{C = 1\} \) all have actual effects on \( \{D = 1\} \). However, neither occurrence is necessary and at least \( \{AB = 1\} \) and \( \{C = 1\} \) are both sufficient for \( \{D = 1\} \). From the perspective of \( t \), \( \{C = 1\} \) is identified as the actual, maximally irreducible, cause of \( \{D = 1\} \). Any framework that disregards this distinction between causes and effects by falsely equating the two would either violate causal exclusion, or provide an incomplete causal description of the transition.

### 4.3 High-order occurrences and composition

That an occurrence can affect several elements (high-order effect), and that the cause of an occurrence can involve several elements (high-order cause) is uncontroversial (Woodward, 2010)\(^{14}\) and also applies to the framework presented here: given a particular occurrence \( x_{t-1} \) in the transition \( \mathbf{x} = \{A = 1\}, \{B = 1\}, \{AB = 11\} \), and the whole power set of \( \mathbf{Y} \) as candidate effects of \( x_{t-1} \), we consider the whole power set of \( \mathbf{X} = \{x_{t-1}\} \) as candidate causes of a particular occurrence \( y_t \). This means, for example, that in all cases shown in Fig. 5, either \( x_{t-1} = \{A = 1\}, \{B = 1\}, \{AB = 11\} \) could be an actual cause of \( y_t = \{C/D/E/F = 1\} \). However, across the set of candidate causes/effects the causal exclusion principle applies: each occurrence has at most one actual cause or effect within a transition, whatever its order (the number of elements involved). The causal principles employed here then single out the actual cause as the minimal set of elements whose actual state is most irreducibly determined by the occurrence, measured by \( c_{\max} \).

While the causal exclusion principle applies to the actual cause and effect of a particular occurrence, the causal composition principle applies to the power set of occurrences within the transition. The proposed framework explicitly acknowledges that there may be high-order occurrences, which have actual causes or actual effects that

\(^{12}\) Initially, the objective of probabilistic accounts of causation was to infer causation, or at least causal relevance, from observed distributions. As demonstrated in (Pearl, 2000, 2009), however, \( p(E|C) > p(E) \) is neither necessary nor sufficient to establish a causal relation between \( E \) and \( C \), highlighting the crucial role of interventions, i.e. the do() operator.

\(^{13}\) Note that Judea Pearl initially proposed maximizing the posterior probability \( p(c|e) \) as a means of identifying the best ‘most probable’ explanation for an occurrence \( e \) (Pearl, 1988; Chapter 5). This approach has later been criticized, among others, by Pearl himself (Pearl, 2000; Chapter 7), as it had been formalized in purely probabilistic terms, lacking the notion of system interventions. Moreover, without a notion of irreducibility, as applied in the present framework, explanations based on \( p(c|e) \) tend to include irrelevant variables (Shimony, 1991; Chajewska and Halpern, 1997).

\(^{14}\) Nevertheless, as illustrated in the case of a simple conjunction (Fig. 5B), some regard each individual element \( A \) and \( B \) as a separate cause by itself, a “contributory cause”, or “part of a cause” (Halpern, 2015).
are irreducible to those of their parts. This notion goes against widespread reductionist views that exclusively assign causal powers to first-order occurrences that involve the elementary constituents of physical systems (here individual system elements), typically without providing a formal notion of causal influence. A simple example of a high-order occurrence with an actual effect is provided in Fig. 5C: in this case, the individual elements A and B do not fulfill the most basic criterion for having an effect on the XNOR element \( E = 1 \) as \( \rho_e(x_{t-1}, y_t) = 0 \), whereas the 2nd-order occurrence \( \{ AB = 11 \} \) has the actual effect \( E = 1 \).

Moreover, as illustrated in the conjunction example (Fig. 5B), the existence of causal links specified by first-order occurrences, in this case \( \{ A = 1 \} \rightarrow \{ D = 1 \} \) and \( \{ B = 1 \} \rightarrow \{ D = 1 \} \), does not exclude the 2nd-order occurrence \( \{ AB = 11 \} \) from having an additional, irreducible effect on \( D = 1 \). Specifically, while \( \{ A = 1 \} \) and \( \{ B = 1 \} \) both constrain the AND-gate D in the same way, the occurrence \( \{ AB = 11 \} \) further raises the probability of \( D = 1 \) compared to the effect of each individual input. Note that, if instead of taking a causal approach, one takes a strictly informational approach focused on predicting \( y_t \) from \( x_{t-1} \) (in an unperturbed system), one might be tempted to disregard such compositional occurrences and their actual effects, since they do not add predictive power. Thus, the actual effect of \( \{ AB = 11 \} \) in the conjunction example of Fig. 5B is informationally redundant, since \( D = 1 \) can be inferred (predicted) from \( \{ A = 1 \} \) and \( \{ B = 1 \} \) alone. However, it is the irreducible mechanistic constraints, and not predictive powers, that are relevant for distinguishing “what caused what” in the various transitions shown in Fig. 5 and that provide an explanation for “what happened”.

The XNOR example can be extended to a “double-biconditional” in transition \( x_{t-1} = \{ ABC = 111 \} \rightarrow y_t \), to illustrate the importance of the composition of occurrences with respect to both actual causes and actual effects (Fig. 11). In the figure, both D and E are XNOR elements that share one of their inputs (element B), and \( \{ AB = 11 \} \leftrightarrow \{ D = 1 \} \) and \( \{ BC = 11 \} \leftrightarrow \{ E = 1 \} \). As illustrated by the cause-repertoire shown in Fig. 11B, and in accordance with D’s and E’s logic function (mechanism), the actual cause of \( D = 1 \) can be described as the fact that A and B were the same, and the actual cause of \( E = 1 \) as the fact that B and C were the same. In addition to these 1st-order occurrences, also the 2nd-order occurrence \( \{ DE = 11 \} \) has an actual cause \( \{ ABC = 111 \} \), which can be described as the fact that all three elements A, B, and C were in the same state. Crucially, this fact is not captured by either the actual cause of \( D = 1 \), or by the actual cause of \( E = 1 \), but only by the irreducible constraints of the 2nd-order occurrence \( \{ DE = 11 \} \). On the other hand, the causal link \( \{ ABC = 111 \} \rightarrow \{ DE = 11 \} \) cannot capture the fact that \( \{ AB = 11 \} \) was the actual cause of \( D = 1 \) and \( \{ BC = 11 \} \) was the actual cause of \( E = 1 \). Of note, in this example the same reasoning applies to the composition of high-order occurrences at \( t - 1 \) and their actual effects.

In sum, irreducible high-order occurrences capture actual, joint mechanistic dependencies between the occurrence’s elements that are not revealed by the actual causes and effects of their parts. Moreover, a high-order occurrence does not exclude lower-order occurrences over their parts, which specify their own actual causes and effects. In this way, high-order and 1st-order occurrences all contribute to the explanatory power of the causal account by the composition principle.

**Fig. 11**: High-order occurrences and composition. (A) Double-Biconditional: Transition and causal account. (B) Cause repertoires corresponding to the two 1st-order and one 2nd-order occurrences with actual causes (see main text).

### 4.4 Irreducibility

In the present framework, having a positive effect ratio \( \rho_e(x_{t-1}, y_t) > 0 \) is necessary but not sufficient for \( x_{t-1} \rightarrow y_t \), and the same for a positive cause ratio \( \rho_c(x_{t-1}, y_t) > 0 \).

As discussed above, high-order occurrences can have actual causes and effects, but only if they are irreducible to their parts. This is illustrated in Fig. 12, in which a transition equivalent to our initial example in Fig. 4 (Fig. 15) Occurrences \( c \) that lower the probability of a subsequent occurrence \( e \) have been termed “preventive causes” (Korb et al., 2011). Rather than counting a negative effect ratio \( \rho_e(c, e) < 0 \) as indicating a possible “preventive effect”, we take the stance that such occurrences \( c \) have no effect at all within the transition of interest, since they actually predict other occurrences \( \bar{E} = \neg e \) that did not happen. By the same logic, a negative cause ratio \( \rho_c(c, e) < 0 \) means that \( c \) is no cause of \( e \) within the transition. Nevertheless, the current framework can in principle quantify the strength of possible “preventive” causes and effects.
12A) is compared against a similar, but reducible transition (Fig. 12C). The two situations differ mechanistically: the OR and AND gate in Fig. 12A receive common inputs from the same two elements, while the OR and AND in Fig. 12C have independent sets of inputs. Nevertheless, the actual causes and effects of all single element occurrences are identical in the two cases. In both transitions, \( \{O = 1\} \) is caused by its one input element in state ‘1’, and \( \{A N D = 0\} \) is caused by its one input element in state ‘0’. What distinguishes the two causal accounts is the additional causal link in Fig. 12A, between the 2nd-order occurrence \( \{O R, A N D = 10\} \) and its actual cause \( \{A B = 10\} \). While \( \{O R, A N D = 10\} \) raises the probability of \( \{A B = 10\} \) (Fig. 12A) and \( \{A D = 10\} \) (Fig. 12C) above chance (\( \pi = 0.25, p_c(x_{t-1}, y_t) > 0 \)), but only \( \{A B = 10\} \) \( \rightarrow \) \( \{O R, A N D = 10\} \) in Fig. 12A is irreducible to its parts. This is shown by partitioning across the MIP (\( \pi = 0.5 \) vs. \( \pi_{M I P} = 0.44 \)) with \( d_c(x_{t-1}, y_t) = 0.17 \). This 2nd-order occurrence thus specifies that the OR and AND gate in Fig. 12A receive common inputs from the same elements.

The measure \( A(x_{t-1} < y_t) \) moreover quantifies the extent to which the causal account \( C \) of transition \( x_{t-1} < y_t \) is irreducible. Thus, \( A(x_{t-1} < y_t) = 0 \) indicates that \( x_{t-1} < y_t \) can either be decomposed into multiple transitions without causal links between them (Fig. 12C,D), or includes irrelevant elements that should be regarded as background conditions (see Fig. 6 and 8) in a more parsimonious causal account (Fig. S1).

4.5 Limitations and possible extensions of actual causation in physical causal models

The literature on actual causation is often rooted in examples involving situational narratives, such as “Billy and Suzy throw rocks at a bottle” (Pearl, 2000; Halpern, 2016), which are then compressed into a causal model to be investigated. Such narratives can serve as intuition pumps, but can also lead to confusion if important aspects of the story are omitted in the causal model fit to the example (Hitchcock, 2007; Paul and Hall, 2013). Issues regarding causation by omission, for instance, are hard to resolve using binary causal models in which the ‘0’ and ‘1’ states are merely labels that could be replaced by ‘a’ and ‘b’. Similarly, examples of preemption (temporal overdetermination) are often represented by causal models that do not explicitly take time into account. In such cases, a helpful strategy is to add abstract variables to the underlying causal model in order to disambiguate a previously inconclusive or unintuitive result (Halpern, 2015).

The present approach aims not so much at capturing common intuitions about causal judgments (Halpern, 2016), but rather at providing a rigorous framework for assessing irreducible causal strength and actual causation (“what caused what”) given a fully specified (complete) model of a physical systems of interacting elements. This puts the explanatory burden on the formal framework of actual causation, rather than on the adequacy of the model. Time is explicitly taken into account, as system elements are synchronously updated at every time step. The framework is applicable to a wide range of systems, deterministic and probabilistic, with feedforward and recurrent architectures, as long as they can be represented by a set of transition probabilities as specified in Eqn. 1. The explanatory power of the proposed framework was illustrated in several examples, which included paradigmatic problem cases involving overdetermination and prevention.

The examples examined in this study have been purposely limited to direct transitions \( x_{t-1} < y_t \) across a single system update, involving binary variables. Abstract examples that raise issues of “causation by omission” or “preemption” have no direct equivalent in these basic types of physical causal models, and may even be physically impossible. Nevertheless, the present framework can be extended in principle to any discrete dynamical system and across multiple time steps by specifying the appropriate
(macro) transition probabilities in Eqn. 1 (Marshall et al., 2016a). An important area in the causation literature concerns causal judgments, which are often tied to notions of causal responsibility and blame assignment. The irreducible effect ratio $\alpha_e$ permits the comparison of the relative effects of different occurrences, and the irreducible cause ratio $\alpha_c$ can measure the relative contributions of different sets of elements to the actual cause of an occurrence. However, addressing the question “who caused what?”, as opposed to “what caused what”, implies modeling an agent with intrinsic causal power and intention (Datta et al., 2015). Future work will therefore combine the present mechanistic framework for actual causation with a mechanistic account of what it means to be an autonomous, causal agent, based on the same set of principles (Marshall et al., in press; Oizumi et al., 2014).

Acknowledgements

We thank Matteo Mainetti for early discussions concerning the extension of IIT to actual causation. This work has been supported by the Templeton World Charities Foundation (Grant #TWCF0067/AB41). L.A. receives funding from the Templeton World Charities Foundation (Grant #TWCF0196).

References:

Albantakis L, Tononi G (2015) The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—from Elementary Cellular Automata to Adapting Animats. Entropy 17:5472–5502.

Ay N, Polani D (2008) Information Flows in Causal Networks. Adv Complex Syst 11:17–41.

Balduzzi D, Tononi G (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Comput Biol 4:e1000091.

Chajewska U, Halpern J (1997) Defining Explanation in Probabilistic Systems. In: Uncertainty in Artificial Intelligence 13 (Geiger D, Shenoy P, eds), pp 62–71. San Francisco, CA: Morgan Kaufmann.

Datta A, Garg D, Kaynar D, Sharma D, Sinha A (2015) Program Actions as Actual Causes: A Building Block for Accountability. In: Proceedings of the Computer Security Foundations Workshop, pp 261–275. IEEE.

Eells E, Sober E (1983) Probabilistic causality and the question of transitivity. Philos Sci 50:35–57.

Glennan S (2011) Singular and General Causal Relations: A Mechanist Perspective. Causality Sci:789.

Good IJ (1961) A Causal Calculus I. Br J Philos Sci 11:305–318.

Halpern J (2015) A Modification of the Halpern-Pearl Definition of Causality. arXiv 1505.00162.

Halpern JY (2016) Actual Causality. Cambridge, MA: MIT Press.

Halpern JYJ, Pearl J (2005) Causes and explanations: A structural-model approach. Part I: Causes. Br J Philos Sci 56:843–887.

Hitchcock C (2007) Prevention, Preemption, and the Principle of Sufficient Reason. Philos Rev 116:495–532.

Hoel EP, Albantakis L, Tononi G (2013) Quantifying causal emergence shows that macro can beat micro. PNAS 110:19790–19795.

Hopkins M, Pearl J (2003) Clarifying the Usage of Structural Models for Commonsense Causal Reasoning. In: Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, pp 83–89. Menlo Park, CA: AAAI Press.

Illari M, Phyllis FR, Williamson J eds. (2011) Causality in the Sciences. Oxford: Oxford University Press.

Janzing D, Balduzzi D, Grosse-Wentrup M, Schölkopf B (2013) Quantifying causal influences. Ann Stat 41:2324–2358.

Korb KB, Nyberg EP, Hope L (2011) A new causal power theory. In: Causality in the Sciences. Oxford: Oxford University Press.

Lewis D (1973) Counterfactuals. Cambridge, MA: Harvard University Press.

Lewis D (1986) Philosophical Papers, volume II. Oxford University Press, Oxford.

Marshall W, Albantakis L, Tononi G (2016a) Black-boxing and cause-effect power. arXiv 1608.03461.

Marshall W, Gomez-Ramirez J, Tononi G (2016b) Integrated Information and State Differentiation. Front Psychol 7:926.

Marshall W, Kim H, Walker SI, Tononi G, Albantakis L (in press) How causal analysis can reveal autonomy in models of biological systems. Philos Trans R Soc A.

Oizumi M, Albantakis L, Tononi G (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Comput Biol 10:e1003588.

Paul LA, Hall EJ (2013) Causation: A user’s guide. Oxford, UK: Oxford University Press.

Pearl J (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers.

Pearl J (2000) Causality: models, reasoning and inference. Cambridge Univ Press.

Pearl J (2009) The Structural Theory of Causations. In: Causality in the Sciences. Oxford: Oxford University Press.
Shimony SE (1991) Explanation, irrelevance and statistical independence. In: Proceedings of the ninth National conference on Artificial intelligence-Volume 1, pp 482–487. AAAI Press.

Suppes P (1970) A Probabilistic Theory of Causality. Amsterdam: North-Holland Pub. Co.

Tononi G (2015) Integrated information theory. Scholarpedia 10:4164.

Tononi G, Boly M, Massimini M, Koch C (2016) Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci 17:450–461.

Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. Proc Natl Acad Sci U S A 96:3257–3262.

Woodward J (2004) Counterfactuals and causal explanation. Int Stud Philos Sci 18:41–72.

Woodward J (2010) Causation in biology: stability, specificity, and the choice of levels of explanation. Biol Philos 25:287–318.

Yablo S (2004) Advertisement for a Sketch of an Outline of a Prototheory of Causation. In: Causation and Counterfactuals (Hall N, Paul LA, Collins J, eds), pp 119–137. Cambridge, MA: MIT Press.
Supplementary Material

Supplementary Fig. S1:

Fig. S1: Linear threshold unit with background condition. Transition and causal account of the linear threshold unit of Fig. 5 (main text) with four inputs and threshold \( k = 3 \), now with \( \{D = 0\} \) as a fixed background condition. In this case, the causal account is irreducible with higher \( \alpha^\text{max} \) values.

Supplementary Fig. S2:

Fig. S2: Disjunction of two conjunctions \((A \land B) \lor C\), extended. (A) Background condition. Transition and causal account of the example in Fig. 6 with \( \{B = 0\} \) as a fixed background condition. The causal account of this transition is still reducible. This is because, given \( \{B = 0\}, \{A = 1\} \) does not have an actual effect on \( \{D = 1\} \) (and, as in Fig. 6, is also not its actual cause). (B) Asymmetric overdetermination. Shown is the causal account of the same system as shown in Fig. 6, but for transition \( \{ABC = 111\} < \{D = 1\} \). The actual cause of \( \{D = 1\} \) is \( \{C = 1\} \), the smallest sufficient condition. Note also that under this transition the causal account is irreducible.
Supplementary Proof 1

The first theorem describes the actual causes and effects for an observation of a linear threshold unit (LTU) \( Y_t \) with \( n \) inputs and threshold \( k \), and its inputs \( X_{t-1} \). First, a series of lemmas are demonstrated based on the transition probabilities \( q_{c,j} \): if \( X_{t-1} = x_{t-1} \) such that the size of the occurrence is \( |x_{t-1}| = c \) and \( j \) of the \( c \) elements in \( X_{t-1} \) are in the ‘ON’ state \( (\sum_{x \in x_{t-1}} x = j) \), then

\[
q_{c,j} = p(Y_t = 1|X_{t-1} = x_{t-1}) = \begin{cases} 
\frac{1}{2n-c} \sum_{i=k-j}^{n-c} \binom{n-c}{i} & \text{if } j \leq k \text{ and } j > k - (n-c) \\
1 & \text{if } j \geq k \\
0 & \text{if } j < k - (n-c)
\end{cases}
\]

First we demonstrate that the probabilities \( q_{c,j} \) are non-decreasing as the number of ‘ON’ inputs \( j \) increases for a fixed size of occurrence \( c \), and that there is a specific range of values of \( j \) and \( c \) such that the probabilities are strictly increasing.

**Lemma 1.1.** \( q_{c,j} \geq q_{c,j-1} \) with \( q_{c,j} = q_{c,j-1} \) iff \( j > k \) or \( j < k - (n-c) \).

**Proof.** If \( j > k \) then

\[ q_{c,j} = q_{c,j-1} = 1. \]

If \( j < k - (n-c) \) then

\[ q_{c,j} = q_{c,j-1} = 0. \]

If \( k - (n-c) \leq j \leq k \) then

\[ q_{c,j} = \frac{1}{2n-c} \sum_{i=k-j}^{n-c} \binom{n-c}{i} = \frac{1}{2n-c} \sum_{i=k-j}^{n-c} \binom{n-c}{i} - \frac{1}{2n-c} = q_{c,j-1} + \frac{1}{2n-c} > q_{c,j-1} \]

Next we demonstrate two results relating the transition probabilities between occurrences of different sizes,

**Lemma 1.2.** \( q_{c,j} = \frac{1}{2} (q_{c+1,j} + q_{c+1,j+1}) \) for \( 1 \leq c < n \) and \( 0 \leq j \leq c \).

**Proof.** If \( j \geq k \) then

\[ q_{c,j} = q_{c+1,j} = q_{c+1,j+1} = 1, \]

so

\[ q_{c,j} = \frac{1}{2} (q_{c+1,j} + q_{c+1,j+1}) = 1. \]

If \( j = k - 1 \) then

\[ q_{c+1,j+1} = 1, \]

and

\[
q_{c,j} = q_{c,k-1} = \frac{1}{2n-c} \sum_{i=1}^{n-c} \binom{n-c}{i} = \frac{1}{2} \left( \frac{1}{2n-c} \sum_{i=1}^{n-(c+1)} \binom{n-(c+1)}{i} + \binom{n-(c+1)}{i} \right) = \frac{1}{2} \left( \frac{1}{2n-c+1} \sum_{i=1}^{n-(c+1)} \binom{n-(c+1)}{i} + \frac{1}{2n-c+1} \sum_{i=0}^{n-(c+1)} \binom{n-(c+1)}{i} \right) = \frac{1}{2} (q_{c+1,j} + 1) = \frac{1}{2} (q_{c+1,j} + q_{c+1,j+1})
\]
If \( j < k - (n - c) \) then
\[
q_{c,j} = q_{c+1,j} = q_{c+1,j+1} = 0,
\]
so
\[
q_{c,j} = \frac{1}{2} (q_{c+1,j} + q_{c+1,j+1}) = 0.
\]

If \( j = k - (n - c) \) then
\[
q_{c+1,j} = 0,
\]
and
\[
q_{c,j} = q_{c,k-(n-c)} = \frac{1}{2^{n-c}} \sum_{i=n-c}^{n-c} \binom{n-c}{i} = \frac{1}{2^{n-c}} \sum_{i=n-(c+1)}^{n-(c+1)} \binom{n-(c+1)}{i} = \frac{1}{2} q_{c+1,j+1} = \frac{1}{2} (q_{c+1,j+1} + q_{c+1,j}).
\]

Finally, if \( k - (n - c) + 1 < j < k - 1 \) then
\[
q_{c,j} = \frac{1}{2^{n-c}} \sum_{i=k-j}^{n-c} \binom{n-c}{i}
\]
\[
= \frac{1}{2^{n-c}} \left( 1 + \sum_{i=k-j}^{n-(c+1)} \binom{n-c}{i} \right)
\]
\[
= \frac{1}{2^{n-c}} \left( 1 + \sum_{i=k-j}^{n-(c+1)} \binom{n-(c+1)}{i} + \sum_{i=k-j}^{n-(c+1)} \binom{n-(c+1)}{i-1} \right)
\]
\[
= \frac{1}{2^{n-c}} \left( \sum_{i=k-j}^{n-(c+1)} \binom{n-(c+1)}{i} + \left( 1 + \sum_{i=k-j}^{n-(c+1)} \binom{n-(c+1)}{i-1} \right) \right)
\]
\[
= \frac{1}{2^{n-c}} \left( \sum_{i=k-j}^{n-(c+1)} \binom{n-(c+1)}{i} + \sum_{i=k-(j+1)}^{n-(c+1)} \binom{n-(c+1)}{i} \right)
\]
\[
= \frac{1}{2} (q_{c+1,j} + q_{c+1,j+1})
\]

\[\square\]

**Lemma 1.3.** If \( c < k \) then \( q_{c,c} < q_{c+1,c+1} \)

*Proof.*
\[
q_{c,c} = \frac{1}{2} (q_{c+1,c} + q_{c+1,c+1}) \quad \text{(Lemma 1.2)}
\]
\[
< \frac{1}{2} (q_{c+1,c+1} + q_{c+1,c+1}) \quad \text{(Lemma 1.1)}
\]
\[
= q_{c+1,c+1}
\]
\[\square\]

Finally, we consider a quantity \( Q(c) \), the sum of \( q \) over all possible states for an occurrence of size \( c \). The value \( Q(c) \) acts as a normalization term when calculating the cause repertoire of occurrence \( \{Y_t = 1\} \). Here we demonstrate a relationship between these normalization terms across occurrences of different sizes.

**Lemma 1.4.** Let \( Q(c) = \sum_{j=0}^{c} \binom{c}{j} q_{c,j} \) then \( Q(c) = \frac{1}{2} Q(c + 1) \)
Proof.

\[ Q(c) = \sum_{j=0}^{c} \binom{c}{j} q_{c,j} \]

\[ = \frac{1}{2} \sum_{j=0}^{c} \binom{c}{j} (q_{c+1,j} + q_{c+1,j+1}) \]

\[ = \frac{1}{2} \left( \sum_{j=1}^{c} \binom{c}{j-1} q_{c+1,j} + \sum_{j=0}^{c} \binom{c}{j} q_{c+1,j} \right) \]

\[ = \frac{1}{2} \left( q_{c+1,c+1} + q_{c+1,0} + \sum_{j=1}^{c+1} \binom{c+1}{j} q_{c+1,j+1} \right) \]

\[ = \frac{1}{2} \left( \sum_{j=0}^{c+1} \binom{c+1}{j} q_{c+1,j+1} \right) \]

\[ = \frac{1}{2} Q(c + 1) \]

□

Using the above lemmas, we are now in a position to prove the actual causes and actual effects in the causal account of a single LTU in the ‘ON’ state. The causal account for a LTU in the ‘OFF’ state follows by symmetry.

**Theorem 1.5.** For a transition \( x_{t-1} \rightarrow y_t \), where \( Y_t \) is a single LTU with \( n \) inputs and threshold \( k \), and \( X_{t-1} \) are its \( n \) inputs, the following holds:

1. The actual cause of \( \{ Y_t = 1 \} \) is \( \{ X_{t-1} = x_{t-1} \} \) where \( |x_{t-1}| = k \) and \( \min(x_{t-1}) = 1 \). Furthermore, the causal strength of the link is

\[ \alpha_{c,\max}(y_t) = k - \log_2 \left( \sum_{j=0}^{k} q_{k,j} \right) \]

2. If \( \min(x_{t-1}) = 1 \) and \( |x_{t-1}| \leq k \) then the actual effect of \( \{ X_{t-1} = x_{t-1} \} \) is \( \{ Y_t = 1 \} \) with causal strength

\[ \alpha_{c}(x_{t-1}, y_t) = \log_2 \left( \frac{q_{c,c}}{q_{c-1,c-1}} \right) > 0, \]

otherwise \( \{ X_{t-1} = x_{t-1} \} \) is reducible.

**Proof.** **Part 1:** Consider an occurrence \( \{ X_{t-1} = x_{t-1} \} \) such that \( |x_{t-1}| = c \leq n \) and \( \sum_{x \in x_{t-1}} x = j \), the probability of \( x_{t-1} \) in the cause-repertoire of \( y_t \) is

\[ \pi(x_{t-1}|y_t) = \frac{q_{c,j}}{Q(c)}. \]

Since \( Y_t \) is a first-order occurrence, there is only one possible partition, and the causal strength of a potential link is thus

\[ \alpha_{c}(x_{t-1}, y_t) = \rho_{c}(x_{t-1}, y_t) = \log_2 \left( \frac{\pi(x_{t-1}|y_t)}{\pi(x_{t-1})} \right) = \log_2 \left( \frac{2^c q_{c,j}}{Q(c)} \right). \]

For a fixed value of \( c \), the maximum value of causal strength occurs at \( j = c \) (since adding ‘ON’ elements can only increase \( q(c,j) \), Lemma 1.1),
Applying Lemma 1.3 and Lemma 1.4, we see that across different values of \(c\), this maximum is increasing for \(0 < c < k\),

\[
\max_{|x_{t-1}|=c} \alpha_c(x_{t-1}, y_t) - \max_{|x_{t-1}|=c} \alpha_c(x_{t-1}, y_t) = \log_2 \left( \frac{2^c q_{c,c}}{Q(c)} \right) - \log_2 \left( \frac{2^c q_{c,c}}{Q(c)} \right) = \log_2 \left( \frac{q_{c+1,c+1}Q(c)}{2^c q_{c,c}Q(c+1)} \right) = \log_2 \left( \frac{q_{c+1,c+1}}{q_{c,c}} \right) > 0,
\]

and that for \(k \leq c\) the causal strength is constant,

\[
\max_{|x_{t-1}|=c+1} \alpha_c(x_{t-1}, y_t) - \max_{|x_{t-1}|=c} \alpha_c(x_{t-1}, y_t) = \log_2 \left( \frac{2^c q_{c+1,c+1}Q(c+1)}{Q(c)} \right) = \log_2 \left( \frac{q_{c+1,c+1}}{q_{c,c}} \right) = \log_2 \left( \frac{1}{1} \right) = 0.
\]

Therefore, the maximum causal strength \(\alpha_c^{\text{max}}(y_t)\) can only occur if \(c \geq k\). We then find the actual cause of \(Y_t = 1\) by minimizing \(|x_{t-1}| = c\) subject to \(|x_{t-1}| = c\) such that \(\alpha_c(x_{t-1}, y_t) = \alpha_c^{\text{max}}(y_t)\). Since \(c \geq k\), we have \(j = c = k\), the actual cause is any set of \(c = k\) elements \(X_{t-1}\) such that \(|x_{t-1}| = c\) and \(\sum_{x \in x_{t-1}} x = c\) (an occurrence of size \(c\) where each element is ‘ON’). Furthermore, when \(c = j = k\),

\[
\alpha_c^{\text{max}}(y_t) = \log_2 \left( \frac{2^k q_{k,k}}{Q(k)} \right) = -\log_2 \left( \sum_{j=0}^{k} q_{k,j} \right) > 0.
\]

**Part 2:** Again, consider occurrences \(X_{t-1} = x_{t-1}\) with \(|x_{t-1}| = c\) and \(\sum_{x \in x_{t-1}} x = j\). The probability of \(y_t\) in the effect repertoire of such an occurrence is

\[
\pi(y_t|x_{t-1}) = q_{c,j} = \begin{cases} 
\sum_{i=k-j}^{n-c} (-1)^{i-c} \binom{n-c}{i} & \text{if } j \leq k \text{ and } j > k - (n-c) \\
1 & \text{if } j \geq k \\
0 & \text{if } j < k - (n-c)
\end{cases}
\]

First, if \(j < c\), then \(\exists x = 0 \in x_{t-1}\) and we can define a partition \(P = \{x_{t-1} - x, y_t\} \cup \{x, \emptyset\}\) such that

\[
\pi_P(y_t|x_{t-1}) = \pi(y_t|(x_{t-1} - x)) \times \pi(\emptyset|x) = \pi(y_t|(x_{t-1} - x)) = q_{c-1,j}
\]

and

\[
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t|x_{t-1})}{\pi_P(y_t|x_{t-1})} \right) = \log_2 \left( \frac{q_{c,j}}{q_{c-1,j}} \right) \leq 0 \quad \text{(Lemma 1.1/1.2),}
\]

so \(x_{t-1}\) is reducible. Next we consider the case where \(j = c\) but \(c > k\). In this case we define a partition \(P = \{(x_{t-1} - x, y_t) \cup \{x, \emptyset\}\} \) (where \(x \in x_{t-1}\) is any element), such that
\[ \pi_P(y_t | x_{t-1}) = \pi(y_t | (x_{t-1} - x)) \times \pi(\emptyset | x) = \pi(y_t | (x_{t-1} - x)) = q_{c-1, e-1}, \]
and since \( c > k \),
\[ \alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t | x_{t-1})}{\pi_P(y_t | x_{t-1})} \right) = \log_2 \left( \frac{q_{c,e}}{q_{c-1,e-1}} \right) = \log_2 \left( \frac{1}{2} \right) = 0, \]
so \( x_{t-1} \) is again reducible. Lastly, we show that for \( j = c \) and \( c \leq k \), that \( x_{t-1} \) is irreducible with actual effect \( \{ Y_t = 1 \} \). All possible partitions of the pair of occurrences can be formulated as \( P = \{ \{ x_{t-1} - x \}, y_t \}, \{ x, \emptyset \} \} \) (where \( x \subseteq x_{t-1} \) with \( |x| = d > 0 \), such that
\[ \pi_P(y_t | x_{t-1}) = \pi(y_t | (x_{t-1} - x)) \times \pi(\emptyset | x) = \pi(y_t | (x_{t-1} - x)) = q_{c-d,c-d}, \]
and
\[ \alpha_c(x_{t-1}, y_t) = \min \log_2 \left( \frac{\pi(y_t | x_{t-1})}{\pi_P(y_t | x_{t-1})} \right) = \min \log_2 \left( \frac{q_{c,e}}{q_{c-d,c-d}} \right). \]
The minimum information partition occurs when \( d = 1 \) (Lemma 1.3) and thus \( \{ X_{t-1} = x_{t-1} \} \) is irreducible with actual effect \( \{ Y_t = 1 \} \) and causal strength
\[ \alpha_c(x_{t-1}, y_t) = \log_2 \left( \frac{q_{c,e}}{q_{c-1,e-1}} \right). \]

**Supplementary Proof 2**

The second theorem describes the actual causes and effects for an observation of a disjunction of conjunctions (DOC) \( Y_t \) that is a disjunction of \( k \) conjunctions, each over \( n_j \) elements, and its inputs \( X_{t-1} \). The total number of inputs to the DOC element is \( n = \sum_{j=1}^{k} n_j \). We consider occurrences \( x_{t-1} \) that contain \( c_j \leq n_j \) elements from each of the \( k \) conjunctions, and the total number of elements is \( |x_{t-1}| = c = \sum_{j=1}^{k} c_j \). To simplify notation, we further define \( \bar{x}_{j,t-1} = \{ x_{i,j,t-1} \}_{i=1}^{n_j} \), an occurrence with \( c_j = n_j \) and \( c_{j'} = 0 \) if \( j' \neq j \). In other words, \( \bar{x}_{j,t-1} \) is the set of elements that make up the \( j^{th} \) conjunction. First, a series of lemmas are demonstrated based on the transition probabilities \( q(s) \):
\[ q(s) = p(Y_t = 1 | X_{t-1} = s) \]
To isolate the specific conjunctions, we define \( s_j \subset x_{t-1} \) to be the state of \( X_{t-1} \) within the \( j^{th} \) conjunction, and \( \bar{s}_j = \bigcup_{i=1}^{j} s_i \subseteq x_{t-1} \) be the state \( X_{t-1} \) within the first \( j \) conjunctions. For a DOC with \( w \) conjunctions, we consider occurrences with \( c_j \) elements from each conjunction, \( X_{t-1} = \{ \{ x_{i,j,t-1} \}_{i=1}^{c_j} \}_{j=1}^{k} \). In the specific case of a disjunction of two conjunctions,
\[ q(s_1, s_2) = \begin{cases} 0 & \text{if } \min(s_1) = \min(s_2) = 0 \\ \frac{1}{2^{\min(s_1) + \min(s_2)}} & \text{if } \min(s_1) = 1, \min(s_2) = 0 \\ \frac{1}{2^{\min(s_1) + \min(s_2)}} & \text{if } \min(s_1) = 0, \min(s_2) = 1 \\ \frac{1}{2^{\min(s_1) + \min(s_2)}} & \text{if } \min(s_1) = \min(s_2) = 1, \end{cases} \]
and in the case of \( k > 2 \) conjunctions, we define the probability recursively
\[ q(s_{k-1}, s_k) = \begin{cases} q(s_{k-1}) & \text{if } \min(s_k) = 0 \\ q(s_{k-1}) + \frac{1-q(s_{k-1})}{2^{\min(s_k)}} & \text{if } \min(s_k) = 1. \end{cases} \]
The first two lemmas demonstrate the effect of adding an additional element to an occurrence. Adding an ‘ON’ input to an occurrence \( x_{t-1} \) can never decrease the probability of \( \{ Y_t = 1 \} \), while adding an ‘OFF’ input to an occurrence \( x_{t-1} \) can never increase the probability of \( \{ Y_t = 1 \} \).
Lemma 2.1. If \( \{x_{t-1} = s\} = \{x'_{t-1} = s', x_{i,j,t-1} = 1\} \), then \( q(s') \leq q(s) \).

Proof. The proof is given by induction. We first consider the case where \( k = 2 \). Assume w.l.g. that the additional element \( x_{i,j,t-1} \) is from the first conjunction \( (c_1 = c'_1 + 1, c_2 = c'_2) \). If \( \min(s'_1) = 0 \) then \( q(s') = q(s) \). If \( \min(s'_2) = 0 \) and \( \min(s'_1) = 1 \) then

\[
\frac{q(s')}{q(s)} = \frac{2^{n_1 - (c'_1 + 1)}}{2^{n_1 - c'_1}} = \frac{1}{2} < 1,
\]

so \( q(s') < q(s) \). Lastly, if \( \min(s'_1) = \min(s'_2) = 1 \) then

\[
\frac{q(s')}{q(s)} = \frac{2^{n_1 + n_2 - (c'_1 + 1) - c'_2} (2^{n_1 - c'_1} + 2^{n_2 - c'_2} - 1)}{2^{n_1 + n_2 - c'_1 - c'_2} (2^{n_1 - (c'_1 + 1)} + 2^{n_2 - c'_2} - 1)} = \frac{2^{n_1 - c'_1} + 2^{n_2 - c'_2} - 1}{2^{n_1 - c'_1} + 2(2^{n_2 - c'_2} - 1)} < 1.
\]

Therefore, when \( k = 2 \) we have that \( q(s') \leq q(s) \). Next, we assume the result holds for \( k - 1, q(s'_{k-1}) \leq q(s_{k-1}) \) and demonstrate the result for general \( k \). Again, assume the additional element is from the first conjunction \( (c_1 = c'_1 + 1, c_j = c'_j \text{ for } j > 1) \). If \( \min(s_k) = 0 \) then

\[
\frac{q(s'_{k})}{q(s_k)} = \frac{q(s'_{k-1})}{q(s_{k-1})} \leq 1,
\]

and if \( \min(s_k) = 1 \) then

\[
\frac{q(s'_{k})}{q(s_k)} = \frac{q(s'_{k-1}) + (1 - q(s'_{k-1}))/2^{n_k-c_k}}{q(s_{k-1}) + (1 - q(s_{k-1}))/2^{n_k-c_k}} = \frac{(2^{n_k-c_k} - 1)q(s'_{k-1}) + 1}{(2^{n_k-c_k} - 1)q(s_{k-1}) + 1} \leq 1.
\]

\[\square\]

Lemma 2.2. If \( \{x_{t-1} = s\} = \{x'_{t-1} = s', x_{i,j,t-1} = 0\} \), then \( q(s') \geq q(s) \).

Proof. The proof is given by induction. We first consider the case where \( k = 2 \). Assume w.l.g. that the additional element is from the first conjunction \( (c_1 = c'_1 + 1, c_2 = c'_2) \). If \( \min(s'_1) = 0 \) then \( q(s') = q(s) \). If \( \min(s'_2) = 0 \) and \( \min(s'_1) = 1 \) then

\[
q(s') = \frac{1}{2^{n_1 - c'_1}} > 0 = q(s)
\]

Lastly, if \( \min(s'_1) = \min(s'_2) = 1 \) then

\[
\frac{q(s')}{q(s)} = \frac{2^{n_2 - c'_2} (2^{n_1 - c'_1} + 2^{n_2 - c'_2} - 1)}{2^{n_2 + n_2 - c'_1 - c'_2} - c'_2} = \frac{2^{n_1 - c'_1} + 2^{n_2 - c'_2} - 1}{2^{n_1 - c'_1}} \geq 1.
\]

Therefore, when \( k = 2 \) we have that \( q(s') \geq q(s) \). Next, we assume the result holds for \( k - 1, q(s'_{k-1}) \geq q(s_{k-1}) \) and demonstrate the result for general \( k \). Again, assume the additional element is from the first conjunction \( (c_1 = c'_1 + 1, c_j = c'_j \text{ for } j > 1) \). If \( \min(s_k) = 0 \) then

\[
\frac{q(s'_{k})}{q(s_k)} = \frac{q(s'_{k-1})}{q(s_{k-1})} \leq 1,
\]

and if \( \min(s_k) = 1 \) then

\[
\frac{q(s'_{k})}{q(s_k)} = \frac{q(s'_{k-1}) + (1 - q(s'_{k-1}))/2^{n_k-c_k}}{q(s_{k-1}) + (1 - q(s_{k-1}))/2^{n_k-c_k}} = \frac{(2^{n_k-c_k} - 1)q(s'_{k-1}) + 1}{(2^{n_k-c_k} - 1)q(s_{k-1}) + 1} \geq 1.
\]
Next, we again consider a normalization term $Q(c)$, which is the sum of $q(s)$ over all states of the occurrence. Here we demonstrate the effect on $Q(c)$ of adding an additional element to an occurrence.

Lemma 2.3. For an occurrence $\{X_{t-1} = x_{t-1}\}$ with $|x_{t-1}| = c > 0$, define $Q(c) = \sum q(s)$. Now consider adding a single element to an occurrence, $x'_{t-1} = \{x_{t-1}, x_i, j_t-1\}$, $(x_{i,j_t-1} \notin x_{t-1})$ such that $c'_{j_1} = c_j + 1$ and $c_j' = c_j$ for $j \neq j_1$, so that $c' = c + 1$. Then $\frac{Q(c')}{Q(c)} = 2$.

Proof. The proof is again given by induction. We first consider the case where $k = 2$, \[
Q(c) = \sum_{s} q(s) = \frac{2^{c_1} - 1}{2^{n_2-c_2}} + \frac{2^{c_2} - 1}{2^{n_1-c_1}} + \frac{2^{n_1-c_1} + 2^{n_2-c_2} - 1}{2^{n_1+n_2-c_1-c_2}}.
\]
Assume w.l.g. that the additional element to the first conjunction ($c_1' = c_1 + 1$). Then we have that \[
\frac{Q(c')}{Q(c)} = \frac{2^{n_1+n_2-c_1-c_2} (2^{n_1} + 2^{n_2} - 1)}{2^{n_1+n_2-c_1-c_2} (2^{n_1} + 2^{n_2} - 1)} = \frac{2^{n_1+n_2-c_1-c_2}}{2^{n_1+n_2-(c_1+1)-c_2}} = 2.
\]
Therefore, when $k = 2$ we have that $\frac{Q(c')}{Q(c)} = 2$. Next, we assume the result holds for $k - 1$ and demonstrate the result for general $k$. Using the recursive relationship for $q$, we get \[
Q_k(c) = \sum_{\bar{s}_k} q(\bar{s}_k) = \sum_{\bar{s}_k} \sum_{\bar{s}_{k-1}} q(\bar{s}_{k-1}, \bar{s}_k) = (2^{c_k} - 1) \sum_{\bar{s}_{k-1}} q(\bar{s}_{k-1}) + \sum_{\bar{s}_{k-1}} \left( q(\bar{s}_{k-1}) + \frac{(1 - q(\bar{s}_{k-1}))}{2^{n_k-c_k}} \right) = \frac{(2^{n_k} - 1)Q_{k-1}(c - c_k) + 2^{c - c_k}}{2^{n_k-c_k}},
\]
Again, assume the additional element is from the first conjunction $c_1' = c_1 + 1$, for the ratio we have \[
\frac{Q_k(c')}{Q_k(c)} = \frac{(2^{n_k} - 1)Q_{k-1}(c' - c_k') + 2^{c' - c_k'}}{(2^{n_k} - 1)Q_{k-1}(c - c_k) + 2^{c - c_k}} = \frac{(2^{n_k} - 1)2Q_{k-1}(c - c_k) + 2^{c - c_k} + 1}{(2^{n_k} - 1)Q_{k-1}(c - c_k) + 2^{c - c_k}} = 2 \frac{2^{n_k} - 1)Q_{k-1}(x'_{t-1}) + 2^{c - c_k}}{(2^{n_k} - 1)Q_{k-1}(x_{t-1}) + 2^{c - c_k}} = 2.
\]
The final two Lemmas demonstrate conditions under which the probability of $\{Y_t = 1\}$ is either strictly increasing or strictly decreasing.
Lemma 2.4. If $\min(x_{t-1}) = 1$, $c_j < n_j \forall j$ and $x'_{t-1} \subset x_{t-1}$ then $q(s') < q(s)$.

Proof. The proof is given by induction. We first consider the case where $k = 2$. Assume w.l.g. that $x_{t-1}$ has an additional element in the first conjunction relative to $x'_{t-1}$ ($c_1 = c'_1 + 1$, $c_2 = c'_2$). The result can be applied recursively for differences of more than one element. First, consider the case where $c_2 < n_2$. The proof is given by induction. We first consider the case where $c_2 < n_2$:

$$
q(s) = \frac{2^{n_1 - c_1} + 2^{n_2 - c_2} - 1}{2^{n_1 - c_1} + 2^{n_2 - c_2} - 1} \times \frac{2^{n_1 + n_2 - c'_2} - 1}{2^{n_1 + n_2 - c'_1}}
$$

$$
= \frac{q(s')}{q(s'_k)} = \frac{2 \left( 2^{n_1 - c_1} + 2^{n_2 - c_2} - 1 \right)}{2^{n_1 - c_1 + 1} + 2^{n_2 - c_2} - 1}

> 1 \quad \text{(since } c_2 < n_2)$$

Therefore, when $k = 2$ we have that $q(s') < q(s)$. Next, we assume the result holds for $k - 1$, $q(s'_{k-1}) < q(s_{k-1})$ and demonstrate the result for general $k$. Again, assume that $x_{t-1}$ and $x'_{t-1}$ differ by a single element in the first conjunction ($c_1 = c'_1 + 1$, $c_j = c'_j$ for $j > 1$). Since $\min(s_k) = 1$,

$$
\frac{q(s_k)}{q(s'_k)} = \frac{q(s'_{k-1}) + (1 - q(s'_{k-1}))/2^{n_k - c_k}}{q(s_{k-1}) + (1 - q(s_{k-1}))/2^{n_k - c_k}} = \frac{(2^{n_k - c_k} - 1)q(s'_{k-1}) + 1}{(2^{n_k - c_k} - 1)q(s_{k-1}) + 1}

> 1.
\]

\]

Lemma 2.5. If $\max(x_{t-1}) = 0$, $c_j \leq 1 \forall j$ and $x'_{t-1} \subset x_{t-1}$ then $q(s) < q(s')$.

Proof. The proof is given by induction. We first consider the case where $k = 2$. Assume w.l.g. that $x_{t-1}$ has an additional element in the first conjunction relative to $x'_{t-1}$ ($c_1 = c'_1 + 1 = 1$, $c_2 = c'_2$). The result can be applied recursively for differences of more than one element. First, consider the case where $c_2 = 1$. Then we have

$$
q(s') = \frac{1}{2^{n_2 - c_2}} > 0 = q(s).
$$

Next consider the case where $c_2 = 0$:

$$
q(s') = \frac{2^{n_1} + 2^{n_2} - 1}{2^{n_1 + n_2}} = \frac{1}{2^{n_2}} \left( \frac{2^{n_1} + 2^{n_2} - 1}{2^{n_1}} \right) = q(s) \left( \frac{2^{n_1} + 2^{n_2} - 1}{2^{n_1}} \right) > q(s).
$$

Therefore, when $k = 2$ we have that $q(s) < q(s')$. Next, we assume the result holds for $k - 1$, $q(s_{k-1}) < q(s'_{k-1})$, and demonstrate the result for general $k$. Again, assume that $x_{t-1}$ and $x'_{t-1}$ differ by a single element in the first conjunction ($c_1 = c'_1 + 1$, $c_j = c'_j$ for $j > 1$). Since $\min(s_k) = 0$,

$$
\frac{q(s_k)}{q(s'_{k-1})} = \frac{q(s_{k-1})}{q(s'_{k-1})} < 1.
$$

Using the above Lemmas, we are now in a position to prove the actual causes and actual effects in the causal account of a single DOC and its inputs. We separately consider the case where the DOC is in the ‘ON’ and the ‘OFF’ state.

Theorem 2.6. For a transition $x_{t-1} \rightarrow y_t$, where $Y_t$ is a single DOC element that is a disjunction of $k$ conditions, each of which is a conjunction of $n_j$ inputs, and $X_{t-1} = \{x_{i,t-1}\}_{i=1}^k = \{\{x_{i,j,t-1}\}_{i=1}^k\}_{j=1}^k$ is the set of its $n = \sum_j n_j$ inputs, the following holds:
Therefore, the maximum value of causal strength is
\[ \min(\bar{x}_{j,t-1}) = 1 \]
and if \( \min(\bar{x}_{j,t-1}) = 1 \) then \( n_j \leq n_{j'} \).

(b) The actual effect of \( \{X_{t-1} = x_{t-1}\} \) is \( \{Y_t = 1\} \) if \( \min(x_{t-1}) = 1 \) and \( c_j = n_j \Rightarrow c_{j'} = 0, j \neq j' \), otherwise \( x_{t-1} \) is reducible.

2. If \( y_t = 0 \),

(a) The actual cause of \( \{Y_t = 0\} \) is \( x_{t-1} \subseteq x_{t-1} \) such that \( \max(x_{t-1}) = 0 \) and \( c_j = 1 \forall j \).

(b) If \( \max(x_{t-1}) = 0 \) and \( c_j \leq 1 \forall j \) then the actual effect of \( \{X_{t-1} = x_{t-1}\} \) is \( \{Y_t = 0\} \); otherwise \( x_{t-1} \) is reducible.

**Proof.** Part 1a: The actual cause of \( \{Y_t = 1\} \). For an occurrence \( \{X_{t-1} = x_{t-1}\} \), the probability of \( x_{t-1} \) in the cause repertoire of \( y_t \) is
\[ \pi(x_{t-1} \mid y_t) = \frac{q(s)}{Q(c)} \]
Since \( Y_t \) is a first-order occurrence, there is only one possible partition, and the causal strength of a potential link is thus
\[ \alpha_c(x_{t-1}, y_t) = \log_2 \left( \frac{\pi(x_{t-1} \mid y_t)}{\pi(x_{t-1})} \right) = \log_2 \left( \frac{2^c q(s)}{Q(c)} \right) = \log_2 (Q_1 q(s)) \]
where \( Q_1 = \frac{2^c}{Q(c)} \forall c \) (Lemma 2.3). If we then consider adding a single element to the occurrence \( x'_{t-1} = \{x_{t-1}, x_{i,j,t-1}' \} \) \( x_{i,j,t-1}' \notin x_{t-1} \) then the difference in causal strength is
\[ \alpha_c(x_{t-1}, y_t) - \alpha_c(x'_{t-1}, y_t) = \log_2 \left( \frac{Q_1 q(s)}{Q_1 q(s')} \right) = \log_2 \left( \frac{q(s)}{q(s')} \right) \]
Combining the above with Lemma 2.2, adding an element \( x_{i,j,t-1} = 0 \) to an occurrence cannot increase the causal strength, and thus occurrences that include elements in state ‘OFF’ cannot be the actual cause of \( y_t \). By Lemma 2.1, adding an element \( x_{i,j,t-1} = 1 \) to an occurrence cannot decrease the causal strength. Furthermore, if \( c_j = n_j \) and \( \min(\bar{x}_{j,t-1}) = 1 \), then \( q(s) = 1 \) and
\[ \alpha_c(y_t, x_{t-1}) = \log_2 (Q_1 q(s)) = \log_2 (Q_1) \]
independent of the number of elements in the occurrence from other conjunctions \( c_{j'} \) and their states \( s_{j'} \) \( j' \neq j \). Since the value \( Q_1 \) does not depend on the specific value of \( j \), it must be the case that this is the maximum value of causal strength, \( \alpha^{\text{max}}(y_t) \). Furthermore, if \( c_j < n_j \forall j \) then
\[ \alpha_c(y_t, x_{t-1}) = \log_2 (Q_1 q(s)) < \log_2 (Q_1) \]
Therefore, the maximum value of causal strength is
\[ \log_2 (Q_1) \]
and this value occurs if and only if there exists \( j \) such that \( c_j = n_j \) and \( \min(\bar{x}_{j,t-1}) = 1 \), i.e. the occurrence includes a conjunction whose elements are all ‘ON’. Finally, minimizing \( |X_{t-1}| \) subject to this constraint gives \( X_{t-1} = \bar{x}_{j,t-1} \) such that \( c_j = n_j, \min(\bar{x}_{j,t-1}) = 1 \), and \( \min(\bar{x}_{j',t-1}) = 1 \Rightarrow c_j \leq c_{j'} \).

Part 1b: Again, consider occurrences \( X_{t-1} = x_{t-1} \) with \( c_j \) elements from each of the \( k \) conjunctions. The effect repertoire of a DOC with \( k \) conjunctions over such occurrences is
\[ \pi(y_t \mid x_{t-1} = s) = q(s) \]
First, if there exists $x \in x_{t-1}$ with $x = 0$ then we can define $x'_{t-1}$ such that $x_{t-1} = \{x'_{t-1}, x\}$, and a partition $P = \{(x'_{t-1}, y_t), \{x, \emptyset\}\}$, i.e. cutting away $x$, such that

$$
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) \times \pi(\emptyset \mid x) = \pi(y_t \mid x'_{t-1}) = q(s'),
$$

By Lemma 2.2, $q(s') \geq q(s)$, and thus

$$
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x'_{t-1})} \right) = \log_2 \left( \frac{q(s)}{q(s')} \right) \leq 0,
$$

so $x_{t-1}$ is reducible. Next we consider the case where $\min(x_{t-1}) = 1$, but there exists $j_1, j_2$ such that $c_{j_1} = n_{j_1}$ and $c_{j_2} > 0$. We define $x'_{t-1} = x_{j_1, t-1}$ and a partition $P = \{(x'_{t-1}, y_t), (x_{t-1} \setminus x'_{t-1}), \emptyset\}$, such that

$$
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) \times \pi(\emptyset \mid (x_{t-1} - x'_{t-1})) = \pi(y_t \mid x'_{t-1}) = q(s'),
$$

and thus

$$
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x'_{t-1})} \right) = \log_2 \left( \frac{q(s)}{q(s')} \right) = \log_2 \left( \frac{1}{1} \right) = 0,
$$

so $x_{t-1}$ is again reducible. We now split the irreducible occurrences into two cases. First, we consider $\min(x_{t-1}) = 1$ and all $c_j < n_j$. All possible partitions of the pair of occurrences can be formulated as $P = \{(x'_{t-1}, y_t), ((x_{t-1} \setminus x'_{t-1}), \emptyset)\}$ (where $x'_{t-1} \subset x_{t-1}$), such that

$$
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) \times \pi(\emptyset \mid (x_{t-1} - x'_{t-1})) = \pi(y_t \mid x'_{t-1}) = q(s'),
$$

and by Lemma 4,

$$
\alpha_c(x_{t-1}, y_t) = \min_P \left( \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x'_{t-1})} \right) \right) = \min_P \left( \log_2 \left( \frac{q(s)}{q(s')} \right) \right) > 0.
$$

So $x_{t-1}$ is irreducible, and its actual effect is $\{Y_1 = 1\}$. Next we consider occurrences such that $\min(x_{t-1}) = 1$ and all $c_j = n_j$ and $c_j = 0$ for $j \neq j_1$ (i.e. $x_{t-1} = x_{j_1, t-1}$). All possible partitions of the pair of occurrences can be formulated as $P = \{(x'_{t-1}, y_t), ((x_{t-1} \setminus x'_{t-1}), \emptyset)\}$ (where $x'_{t-1} \subset x_{t-1}$), such that

$$
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) \times \pi(\emptyset \mid (x_{t-1} - x'_{t-1})) = \pi(y_t \mid x'_{t-1}) = q(s'),
$$

$$
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x'_{t-1})} \right) = \log_2 \left( \frac{q(s)}{q(s')} \right) = \log_2 \left( \frac{1}{q(s')} \right) > 0,
$$

and $x_{t-1}$ is again irreducible with actual effect $\{Y_1 = 1\}$.

**Part 2a:** The actual cause of $\{Y_t = 0\}$. For an occurrence $\{X_{t-1} = x_{t-1}\}$ the cause repertoire of $y_t$ is

$$
\pi(x_{t-1} \mid y_t) = \frac{1 - q(s)}{2^\varepsilon - Q(c)}.
$$

Since $Y_t$ is a first-order occurrence, there is only one possible partition, and the causal strength of a potential link is thus

$$
\alpha_c(x_{t-1}, y_t) = \log_2 \left( \frac{\pi(x_{t-1} \mid y_t)}{\pi(x_{t-1})} \right) = \log_2 \left( \frac{2^\varepsilon (1 - q(s))}{2^\varepsilon - Q(c)} \right) = \log_2 \left( \frac{Q_0 q(s)}{q(s')} \right),
$$

where $Q_0 = \frac{x}{2^\varepsilon - Q(c)} \forall c$ (Lemma 2.3). If we then consider adding a single element to the occurrence $x'_{t-1} = \{x_{t-1}, x'_{j_1, t-1}\}$ ($x'_{j_1, t-1} \notin x_{t-1}$), then the difference in causal strength is

$$
\alpha_c(x_{t-1}, y_t) - \alpha_c(x'_{t-1}, y_t) = \log_2 \left( \frac{Q_0 (1 - q(s))}{Q_0 (1 - q(s'))} \right) = \log_2 \left( \frac{1 - q(s)}{1 - q(s')} \right).
$$
By Lemma 2.2, adding an element \( x = 1 \) to an occurrence cannot increase the causal strength, and thus occurrences that include elements in state ‘ON’ cannot be the actual cause of \( y_t \). By Lemma 2.1, adding an element \( x = 0 \) to an occurrence cannot decrease the causal strength. Furthermore, if \( c_j > 0 \) \( \forall j \) and \( \max(x_{t-1}) = 0 \), then

\[
\alpha_c(y_t, x_{t-1}) = \log_2 (Q_0(1 - q(s))) = \log_2 (Q_0),
\]

independent of the actual values of \( c_j \). Since this holds for any set of \( c_j \) that satisfies the conditions, it must be the case that this value is \( \alpha^\text{max}(y_t) \). Furthermore, if there exists \( j \) such that \( c_j = 0 \) then

\[
\alpha_c(y_t, x_{t-1}) = \log_2 (Q_0(1 - q(s))) < \log_2 (Q_0).
\]

Therefore, the maximum value of causal strength is

\[
\log_2 (Q_0),
\]

and this value occurs if and only if \( c_j > 0 \) \( \forall j \) and \( \max(x_{t-1}) = 0 \), i.e. the occurrence contains elements from every conjunction, and only elements whose state is ‘OFF’. Finally, minimizing \( |X_{t-1}| \) subject to this constraint gives any occurrence \( \{X_{t-1} = k \} \) such that \( c_j = 1 \) \( \forall j \) and \( \max(x_{t-1}) = 0 \).

**Part 2b:** Again, consider occurrences \( X_{t-1} = x_{t-1} \) with \( c_j \) elements from each of \( k \) conjunctions. The probability of \( y_t \) in the effect repertoire of \( x_{t-1} \) is

\[
\pi(y_t \mid x_{t-1} = s) = 1 - q(s).
\]

First, if there exists \( x_{i,j,t-1} \in x_{t-1} \) such that \( x_{i,j,t-1} = 1 \) then we can define \( x'_{t-1} \) such that \( x_{t-1} = \{x'_{t-1}, x_{i,j,t-1}\} \) and a partition \( P = \{\{x'_{t-1}, y_t\}, \{x_{i,j,t-1}, \emptyset\}\} \) such that

\[
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) = \pi(y_t \mid x_{t-1} = 1 - q(s')).
\]

By Lemma 2.1, we have \( 1 - q(s) \leq 1 - q(s') \), and thus

\[
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x_{t-1})} \right) = \log_2 \left( \frac{1 - q(s)}{1 - q(s')} \right) \leq 0,
\]

so \( x_{t-1} \) is reducible. Next we consider the case where \( \max(x_{t-1}) = 0 \), but there exists \( j \) such that \( c_j > 1 \). We define \( x'_{t-1} \) such that \( x_{t-1} = \{x'_{t-1}, x_{i,j,t-1}\} \), and a partition \( P = \{\{x'_{t-1}, y_t\}, \{x_{i,j,t-1}, \emptyset\}\} \), such that

\[
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) = \pi(y_t \mid x_{t-1} = 1 - q(s') = 1 \]

and

\[
\alpha_c(x_{t-1}, y_t) \leq \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x_{t-1})} \right) = \log_2 \left( \frac{1 - q(s)}{1 - q(s')} \right) = \log_2 \left( \frac{1}{2} \right) = 0,
\]

so \( x_{t-1} \) is again reducible. Finally, we show that occurrences \( x_{t-1} \) are irreducible if \( \max(x_{t-1}) = 0 \) and all \( c_j \leq 1 \). All possible partitions of the pair of occurrences can be formulated as \( P = \{\{x'_{t-1}, y_t\}, \{(x_{t-1} \setminus x'_{t-1}), \emptyset\}\} \) (where \( x'_{t-1} \subset x_{t-1} \)), such that \( c'_j \leq c_j \forall j \), and \( c' < c \). Then

\[
\pi(y_t \mid x_{t-1}) = \pi(y_t \mid x'_{t-1}) = \pi(y_t \mid (x_{t-1} - x'_{t-1})) = \pi(y_t \mid x_{t-1} = 1 - q(s'),
\]

and by Lemma 2.5,

\[
\alpha_c(x_{t-1}, y_t) = \min_P \left( \log_2 \left( \frac{\pi(y_t \mid x_{t-1})}{\pi(y_t \mid x_{t-1})} \right) \right) = \min_P \left( \log_2 \left( \frac{1 - q(s)}{1 - q(s')} \right) \right) > 0.
\]

So \( \{X_{t-1} = x_{t-1}\} \) is irreducible, and its actual effect is \( \{Y_1 = 1\} \).