Computational Brain & Behavior (2023) 6:377–399
https://doi.org/10.1007/s42113-022-00156-z

ORIGINAL PAPER

Discovering Common Hidden Causes in Sequences of Events

Simon Valentin[1] · Neil R. Bramley[2] · Christopher G. Lucas[1]

Accepted: 9 September 2022 / Published online: 11 November 2022

Abstract

Human cognition is marked by its ability to explain patterns in the world in terms of variables and regularities that are not directly observable, e.g., mental states, natural laws, and causal relationships. Previous research has demonstrated a capacity for inferring hidden causes from covariational evidence, as well as the use of temporal information to identify causal relationships among observed variables. Here we explore the human ability to use temporal information to make inferences about hidden causes, causal cycles, and other causal relationships, without relying on interventions. We examine two behavioral experiments and compare participants' judgments to those of Bayesian computational-level models that use temporal order and delay information to infer the causal structure behind observed event sequences. Our results indicate that participants are able to use order and timing information to discover hidden causes, and make inferences about causal structures relating hidden and observable variables. Computational modeling indicates that most participants are best described by normative delay model predictions, but also reveals several clusters of participants who made unexpected inferences, suggesting opportunities to enrich future models of human causal reasoning.

Keywords Causal inference · Causal learning · Event cognition · Hidden variables · Latent variables · Bayesian models

Introduction

People have a remarkable capacity to make sense of the sparse, noisy, and ambiguous stream of data that makes up everyday experience. We infer causal relationships to explain events that occur close in time—such as between pressing an unmarked button on a hotel TV remote and seeing the TV turn on—but also between events much further apart in time—such as eating fast food and having an upset stomach some hours later. Often this involves positing the existence of unobserved (or *latent*) causes as well as the (inherently unobservable) causal relationships. For instance, for the example above, instead of concluding the food caused the symptoms, we may think that both deciding to eat fast food and having an upset stomach are due to stress at work or school. Similarly, sitting on a train and observing that several people are repeatedly picking up their phones at similar times, we may infer that they are reacting to the same messages in a group chat or news notifications.

Coming up with an appropriate generative model of the external environment is valuable for any cognitive agent, allowing for accurate prediction and effective control in pursuit of goals, as well as sub-serving explanation and communication. Where latent variables actually exist, identifying them also tends to result in a better and more compact representation than attempting to do without them (e.g., Gershman & Niv, 2010). Previous studies on causal reasoning have shown that adults and children, even as young as 10 months, can use covariation information to learn about hidden causes (Kushnir et al., 2003, 2010; Saxe et al., 2005; Lucas et al., 2014; Rottman et al., 2011). In particular, Lucas et al. (2014) showed that adults can correctly infer the presence of one or several hidden elements, as well as the functional form of their causal mechanism to explain the behavior of a black-box machine.

The possibility of common hidden causes, or *latent confounders*—i.e., causally relevant variables that have not been or cannot be observed—creates a challenge for the discovery of causal relationships. This is because dependencies between any two events or variables can always be explained by the idea that both are being influenced by some

✉ Simon Valentin
s.valentin@ed.ac.uk

[1] School of Informatics, University of Edinburgh, Edinburgh, UK

[2] Department of Psychology, University of Edinburgh, Edinburgh, UK

unobserved third variable. As a simple example, an observed correlation between X and Y is consistent with a variety of causal models: Perhaps X causally influences Y, perhaps Y causally influences X; or perhaps some unobserved variable H causally influences both X and Y. These are also not mutually exclusive possibilities; it could also be the case that there is a bidirectional causal influence with X influencing Y even as Y influences X. Worse, X and Y could be constituents of some larger feedback loop involving multiple hidden variables.

Causal graphical models (CGMs, also known as causal Bayesian networks; Pearl, 1995, 2009) can help us formalize and better understand this problem.[1] CGMs have become a dominant tool for causal inference both in data science and as a framework for modeling causal cognition (e.g., Koller and Friedman, 2009; Griffiths & Tenenbaum, 2005; Gopnik & Tenenbaum, 2007; Tenenbaum & Griffiths, 2001; Bramley et al., 2017). A CGM represents causal relationships between random variables using directed edges and parameters encoding the causal mechanisms connecting causes with their effects. This formalism has the constraint that the resulting graph must be acyclic, i.e., there can be no path from any node in the graph back to itself, meaning there is no natural way to represent feedback loops. As in the example involving X and Y above, when learning a CGM from observational contingency data, there is a strict upper bound on structural identifiability. Causal structures that have different interventional semantics can be Markov equivalent, meaning they imply identical (conditional) independencies in the absence of interventions and so cannot be distinguished from observational data alone (e.g., Pearl, 2009; Peters et al., 2017; Heinze-Deml et al., 2018).

The gold standard for uniquely determining causal structure is to perform experiments, manipulating causal variables and observing what else changes (Pearl, 2009; Cartwright, 2007). However, interventions are not always practical or even possible; in many settings, they may be unethical or prohibitively costly. While people are capable intuitive experimenters (Gopnik et al., 2007; Cook et al., 2011; Kushnir & Gopnik, 2005; Lagnado & Sloman, 2004; Steyvers et al., 2003), they may also make use of the richness of observational data to tackle the challenging problem of inferring causal structure (e.g., Rothe et al., 2018). Given the abundance of observational data, ignoring such information would severely limit people's ability to make sense of the world around them. Critically, temporal information can provide cues and constraints on causal structure that are missed when only considering "static" contingency information, which may be particularly important for tackling the challenging problem of inferring causal structures involving hidden causes.[2]

People do not generally encounter contingency data directly, but rather events occurring over time, often without the additional information that would allow them to build a contingency table (e.g., about what would constitute an "independent trial"). Temporal order and delay between events have been linked to learning since the early days of psychology, featuring in basic accounts of animal and human learning and conditioning (e.g., Grice, 1948; Michotte, 1946). The connection between order and causality has been noted since the earliest work on causality, for instance by Hume, as causes are assumed to precede their effects (Hume, 1740). Recent research on the role of time in human causal structure learning has shown that people readily use temporal information to infer causal relationships among observed events, making use of both temporal order (Rottman & Keil, 2012; Bramley et al., 2014) and delay information (Bramley et al., 2018; Gong et al., 2022). Underlining the importance of temporal information, people have also been found to make inferences that align with the temporal order of events, even if this temporal information is at odds with covariation evidence (Lagnado & Sloman, 2006; Rottman & Keil, 2012). Regarding more nuanced temporal delay information, it is well established that longer delays between two events lead to weaker judgments of causality, other things being equal (Grice, 1948; Shanks et al., 1989). Explanations for this observation consider working memory capacity constraints (Ahn et al., 1995; Einhorn & Hogarth, 1986), but also the normative rationale that longer delays imply more events may have occurred in the meantime that could explain the effect (Buehner & May, 2003; Lagnado & Speekenbrink, 2014). Shorter delays are not invariably seen as more causal, however: People are also able to adapt their expectations to specific domains. As in the introductory examples, we expect the delay between pressing the power button on a TV controller and seeing the device turn on to be short but between eating fast food and developing symptoms of an upset stomach to be much longer (Garcia et al., 1966; Buehner & McGregor, 2006a). Indeed, violations of the expectations about causal delays have been shown to reduce judgments of causal strength (Greville & Buehner, 2010). Research focusing on the problem of identifying the structure of fully observed and acyclic causal systems suggests that people use order information to rule out incompatible causal structures, but are also able to use the duration

[1] Note that we consider directed CGMs, but there have also been proposals for including undirected edges, see, e.g., Peters et al. (2017) for a discussion.

[2] In some cases, it is possible to rely on auxiliary information and assumptions to go beyond the equivalence classes that can be recovered from observational contingency data, e.g., by making assumptions about the (non-)linearity of causal mechanisms and (non-) Gaussianity of noise distributions (see Eberhardt, 2017; Peters et al., 2017; Guo et al., 2020).

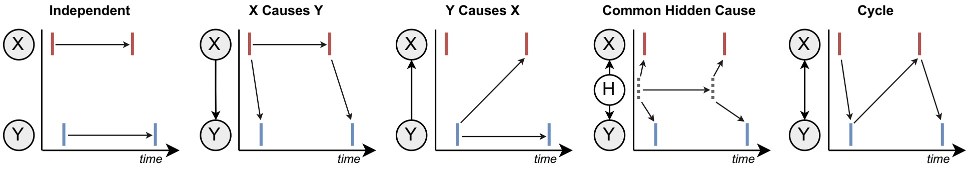


Fig. 1 Causal structure between events under each hypothesis. Events are denoted as red and blue vertical bars, for X and Y events, respectively. Hidden cause events are denoted by dotted vertical gray bars

and variability of causal delays to make more fine-grained judgments (Bramley et al., 2018).

In this work, we build on prior research into how people infer the presence of hidden causes as well as on work that studies the role of temporal information in shaping people's causal inferences. Specifically, we study how people use temporal information to infer the causal structure giving rise to observed sequences of events when this can involve hidden causes and causal cycles (see Fig. 1). Across two experiments covering different domains, we compare people's causal judgments to Bayesian structure learning models that use order or delay information.

Formal Framework

Our approach follows the tradition of rational analysis (Anderson, 1991) and computational-level models (Marr, 1982). That is, we derive normative predictions under different assumptions about how people may construe the learning problem and compare these predictions to human judgments. We take causal structure learning to be a probabilistic inverse problem, a common perspective that has been successful in explaining a wide range of phenomena (e.g., Griffiths & Tenenbaum, 2005; Gopnik & Tenenbaum, 2007). More precisely, we represent structure learning as a problem of Bayesian inference over a hypothesis space of possible causal generative models. This begins with the learner's prior beliefs about the set of possible causal structures S and their parameters θ_s for all $s \in S$, represented as $p(s)$ and $p(\theta_s|s)$, respectively. Given data $\mathcal{D}$, a learner updates their beliefs about causal structures via Bayes' theorem:

$$p(s \mid \mathcal{D}) \propto p(\mathcal{D} \mid s)p(s). \tag{1}$$

Here, $p(\mathcal{D} \mid s)$ is the marginal likelihood of structure s, having integrated out our uncertainty about the parameters and potential hidden causes. However, this marginal is typically not available in closed form, as this involves integrating over unknown parameters, even for structures without hidden causes:

$$p(\mathcal{D} \mid s) = \int p(\mathcal{D} \mid \boldsymbol{\theta}_s, s)p(\boldsymbol{\theta}_s \mid s)d\boldsymbol{\theta}_s. \tag{2}$$

Evaluating marginal likelihood for structures with hidden hidden causes involves even more uncertainty, since the likelihood of parameters $p(\mathcal{D} \mid \boldsymbol{\theta}_s)$ depends on the values of the hidden causes ($\mathbf{h}$), which must also be marginalized over:

$$p(\mathcal{D} \mid s) = \iint p(\mathcal{D} \mid \mathbf{h}, \boldsymbol{\theta}_s, s)p(\mathbf{h} \mid \boldsymbol{\theta}_s, s)p(\boldsymbol{\theta}_s \mid s)d\mathbf{h}d\boldsymbol{\theta}_s. \tag{3}$$

The integrals in these expressions can usually not be solved in closed form, so they require the use of an approximation scheme. We next discuss existing proposals to modeling time in causal relationships, before discussing our approach.

Existing Modeling Approaches

Static CGMs serve as interpretable and compact static representations of the causal relationships between random variables; however, the temporal dynamics between individual events are not typically represented explicitly. The idea of taking into account temporal information for discovering causal relationships is not new, considering, e.g., Granger causality (Granger, 1969), dynamic causal modeling (Friston et al., 2003) and advances in the machine learning literature (e.g., Didelez, 2008; Pamfil et al., 2020; Löwe et al., 2020; Malinsky & Spirtes, 2019; Strobl, 2019; Mastakouri et al., 2021).

A popular variant of graphical models for modeling dynamic relationships or systems that evolve over time is given by dynamic Bayesian networks (DBNs; Dean & Kanazawa, 1989), where—as for for static CGMs—edges can be attached with causal semantics. However, one downside is that DBNs typically require a discretization of continuous time into discrete steps. Continuous time Bayesian networks (CTBNs; Nodelman et al., 2002) extend DBNs to represent structured stochastic processes in continuous time, thereby mitigating the problem of choosing a discretization of time in DBNs. However, both standard DBNs and standard CTBNs implicitly assume that delays between events are memoryless, following exponential distributions (Murphy, 2012). This assumption does not hold for many real-world phenomena, such as incubation periods expressing the delay between exposure to a virus and the first sign of symptoms.

A related approach that relaxes the DAG representation of CGMs is to introduce an undirected edge to capture both common hidden causes and cyclic relationships, as in so called "chain graphs" (Lauritzen & Richardson, 2002). However, this amounts to declaring the distinction between causal cycles and hidden causes as unidentifiable, even though these graphs respond differently to interventions: While intervening on one of the observed variables might cause the other variable to activate in a cycle, such interventions would have no effect in a CHC structure, as there is no direct causal connection between the observed variables. Overall, despite many advancements, challenges remain for inferring causality from temporal as well as atemporal data (e.g., Glymour et al., 2019).

A different class of models that were applied to human causal cognition are based on point processes and represent causes that influence the rate with which other events occur in continuous time (Pacer et al., 2012; Pacer & Griffiths, 2015). Recent modeling work has also addressed the question of how people infer causality between specific event instances. This has been studied by Stephan et al. (2020), who focus on singular causes rather than causal structure over multiple observations with a modeling framework that is related to the delay model we discuss below. A different line of work has focused on the question of how people infer causal structure between continuous causes in continuous time, as opposed to events in continuous time (Davis et al., 2020). Moreover, while the present work focuses on studying human behavior at the computational level, prior research by Fernando (2013) presents an approach to learning causal structure between events at an implementational (i.e., neural) level.

Events in Continuous Time

We are interested in the causal relationships between the onsets of events, which can be treated as points on the real line. This means that two events never occur at exactly the same moment. We also assume that effects never precede their causes. Figure 1 shows an example of how four events imply different sets of cause–effect delays under five different causal structures.

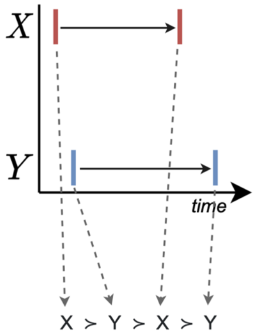


Fig. 2 Parsing of events on the real line for an order-representation, where colored vertical bars represent point events. In this case X and Y are assumed independent (as represented by the causal directed edges between events), but the sequence $X \succ Y \succ X \succ Y$ is possible under all five structures, except for $Y \rightarrow X$

Order Model

We first describe an inference model that is sensitive only to order information and disregards delays between events. For this, we generalize previous work (Rottman and Keil, 2012; Bramley et al., 2014, 2018) in order to accommodate hidden causes and cycles. Our order model has likelihoods that depend only on the order of events in a given sequence, so is agnostic about the exact length of any inter-event delay distributions (see Fig. 2). We construct generative order models for each causal structure, compactly representing each as a probabilistic finite state machine (PFSM; Vidal et al., 2005) (see Appendix 3). Depending on the causal structure, different sequences of events are possible.

The model in which X and Y are independent imposes no restrictions on transitions between events: X might be followed by another occurrence of X or by a Y, and vice versa. For $X \rightarrow Y$ (that is, X causes Y), we assume there is only one possible transition for each state: cause X is invariably followed by its effect Y and X does not reoccur until its previous activation (and causal chain to Y) has run its course, meaning that Y is always followed by X. Essentially, for the order model (but not the delay model, as discussed below), we assume causes are blocked from re-occurrence while they are still involved in producing their effects. An alternative could be to assume that multiple causal influences are able to travel between a cause and its effect simultaneously. We return to this assumption in the discussion. The structure $Y \rightarrow X$ has the mirrored semantics of $X \rightarrow Y$. A common hidden cause can produce activations of X and Y in succession, but uniquely and as opposed to the independent structure, implies that the same variable will never activate more than twice in a row, following the assumption described above. For causal cycles, the order-only model assumes either observable event might initialize the observation sequence, but after this the observables activate in turn.

Following prior work (Bramley et al., 2018), we complete each order structure by setting the transition probability for a state k to $\frac{1}{\text{outdegree}(k)}$, where the outdegree is defined as the number of outgoing edges from a state, implementing the principle of the Bayesian Ockham's razor (e.g., Myung & Pitt, 1997). For example, we assume there is always a $\frac{1}{2}$ chance of X occurring as the next event in the independent structure (see Appendix 3).[3]

Delay Model

For our generative model of events and their delays, we use a variant of a dynamic Bayesian network (DBNs; Dean & Kanazawa, 1989) representation, as a CGM in which nodes denote *when* components of the causal system activate. Edges correspond to parameterized delay distributions controlling the intervals between activations of a cause and effects. Root causes are assumed to cause their own recurrence with their own set of parameters controlling their inter-event distributions. Here we restrict out attention to causal relationships in which each effect event can only have at most one cause event, and effect events appear in the order their cause events occurred in. For instance, for a $X \rightarrow Y$ structure, the sequence $x^{(1)} \succ x^{(2)} \succ y^{(1)} \succ y^{(2)}$ (where $\succ$ denotes precedence with respect to the temporal order) would be consistent with the delay model, whereas the sequence $x^{(1)} \succ x^{(2)} \succ y^{(2)} \succ y^{(1)}$ would not be consistent. This assumption is weaker than the assumption for the order model, as the delay model does not assume that causes are blocked until their effects have occurred, but only assumes that the problem of causal attribution is resolved via temporal precedence; we discuss this point further in the general discussion. To generate data from such a model, one samples root cause activations, then samples causal delays to their effects, unrolling the graph into a tree of event timings. Inference then amounts to "reverse engineering" the causal structure most likely to have given rise to the set of timings observed, taking into account prior beliefs about their plausibility.

Following prior work, we use the gamma distribution to model delays (e.g., Bramley et al., 2018; Gong et al., 2022; Stephan et al., 2020). Gamma distributions have positive, i.e., $\in (0, \infty)$, support and can capture beliefs about the expectation and variance of delay distributions. Gamma distributions are typically defined by a shape k and a scale θ (or alternatively, by a shape α and rate β). As opposed to the exponential distribution (which is a special case of the gamma distribution), the gamma distribution allows for modeling non-memoryless delay distributions. That is, gamma distributions can capture expectations about when an effect will happen after observing a cause as well as how much variability there is around this expectation. For easier interpretability, we express gamma distributions using a standard reparametrization in terms of their mean $\mu = k\theta$ and variance $\sigma^2 = k\theta^2$. For additional background on the gamma distribution, see Appendix 4.

We display static summary graphs along with example event sequences in Fig. 1. For our computational analysis, we index events of each type separately by their i-th occurrence. In the unrolled graphical representation, edges thus represent parameterized gamma delays between occurrences of events (represented as nodes). For instance, if $x^{(i)}$ is connected to $y^{(i)}$ by an edge under a particular structure hypothesis, $x^{(i)}$ occurred at some time $x^{(i)} = 1s$ and we observe a delay of $0.4s$, then the value of $y^{(i)}$ is $y^{(i)} = x^{(i)} + 0.4s = 1.4s$.

For the independent structure, $x^{(i)}$ causes $x^{(i+1)}$ and $y^{(i)}$ causes $y^{(i+1)}$, such that there is no inherent dependence between occurrences of X and the next occurrence of Y. Under $X \rightarrow Y$, each $x^{(i)}$ is caused by $x^{(i-1)}$, while each $y^{(i)}$ is caused by $x^{(i)}$ and thus is independent of $y^{(i-1)}$ conditional on $x^{(i)}$. For the common hidden cause structure, the occurrences of $x^{(i)}$ and $y^{(i)}$ are taken to be caused by activations of a hidden cause $h^{(i)}$, which was self-caused by $h^{(i-1)}$. We further assume causal delays from $h^{(i)}$ to $x^{(i)}$ and $h^{(i)}$ to $y^{(i)}$ have tied parameters, such that occurrences of X neither systematically succeed or precede occurrences of Y. Coupled with our assumptions about delay distributions, this means that common hidden causes entail effects that are close to one another in time, but occur in arbitrary orders. This assumption of tied parameters may be justified in real-world settings whenever the observed variables have a shared causal mechanism, e.g., if both observed variables are instances of the same "type" of variable (as is the case for our cover stories below), but see the discussion about the possibility of relaxing this assumption. Lastly, the causal cycle resembles $X \rightarrow Y$ (or $Y \rightarrow X$) but distinguishes itself by two characteristics, which arise from a shared cause-effect mechanism: (1) delays from X to Y and Y to X are symmetric; hence, the parameters are tied; and (2) a sequence of observations might begin with either X or Y.

Overview of Experiments

We ran two experiments in which participants watched a series of short videos. Participants were allocated to different cover stories, but videos always showed two colored circles on a gray background, corresponding to the two observable components X and Y of some causal scenario of unknown structure. Activations were then visualized by the requisite component flashing briefly. Participants had to

[3] As noted above, since time is continuous and we treat events as points, the chance of two coinciding exactly is zero. However, perception is not infinitely sensitive, representing continuous numbers on a computer requires discretization, and computer screens have a finite refresh rate. Therefore, events occurring very close together may be perceived as simultaneous. In a few cases in our stimuli, two events occurred within 50ms of one another, so potentially appearing to coincide. In these cases, we compute order model likelihoods by marginalizing over the two possible orders, assuming both are equally probable.

watch the videos and make a forced choice from the set of five candidate causal structures.

Experiment 1

Methods

Participants

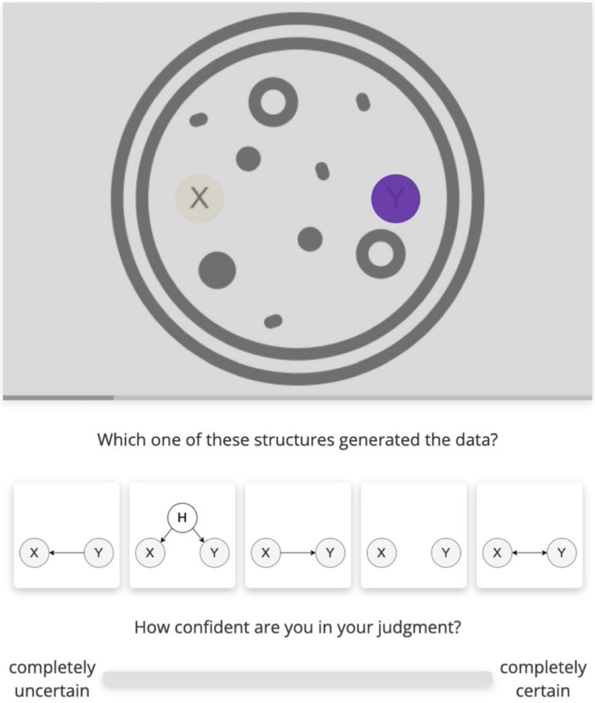
Fifty adults (18 female, mean age 34.30 years, $SD = 10.70$) participated in the experiment in return for a base payment of £1.50 and performance-related bonuses of up to £1.20 resulting in an average compensation of £11.46/h. Participants took, on average, 16.25 ($SD = 9.09$) minutes to complete the task.

The experiment was conducted online with participants recruited via Prolific Academic (www.prolific.co). The experiment was programmed in Scala.js and ran as a standard client-side experiment. In order to ensure high data quality, participants were required to have a 99% completion rate on previous studies on Prolific, as well as between 100 and 10,000 previous submissions. The study was pre-registerd on OSF (https://osf.io/e5d23).

The sample size for our first experiment is based on a power analysis using G*Power (Faul et al., 2009). We consider binomial tests for each of the five conditions, where the stimuli were sampled from the generative structure (that is, excluding the prior elicitation condition). We are interested in detecting deviations from the expected proportion of chance responses (0.2 for 5 response options) for the ground-truth category relative to all other categories. For a power of 0.95, a Bonferroni-adjusted alpha level of $\frac{0.05}{5} = 0.01$ (for running five separate tests), a proportion of 0.2 under the null-hypothesis, and a proportion (to be interpreted as an effect size) of 0.5 under the alternative hypothesis, we obtain a required total sample size of 44. Including a safety margin, we thus obtain a total sample size of 50 participants.

Design and Procedure

Participants were asked to identify the causal relationships between different species of bioluminescent bacteria and told they would be paid £0.20 for each trial in which they identified this correctly. Concretely, their task was to identify the causal structure giving rise to observed sequences of events. We constructed video stimuli that showed two bacteria labeled X and Y which would occasionally light up and presented five possible causal hypotheses in the form of graphical diagrams relating the bacteria to one another and to a potential hidden cause (see Fig. 3 for a screenshot of the interface).



Which one of these structures generated the data?

How confident are you in your judgment?

Fig. 3 Experiment interface, showing the stimulus display (top) and structure judgment, including a confidence slider, for the bacteria cover story (bottom). This frame shows a Y event

Participants were first trained on how to interpret each diagram, see Appendix 1 for the training participants underwent. Participants were also told that in addition to the observed X and Y bacteria, a hidden bacterium might also be present, and that this might influence the occurrence of illuminations of the observed bacteria. Participants were also instructed that some causal structures may never be the correct answer and some might be correct for several of the trials. After completing instructions and two comprehension checks (see Appendix 1), each participant completed six trials in randomized order. Five of these involved watching 35 s videos containing between 12 and 15 occurrences of each bacterium X and bacterium Y lighting up. In reality, each of the five causal structures was used to generate the data for exactly one of these trials. One additional trial (occurring randomly in the sequence of trials) did not include any observed data. Participants were instead instructed: "This recording must have gone missing. Even though there is no video recording available, please give your best guess about which structure it might have been anyway." This trial served to probe participants' prior expectations about the plausibility of the different structures without seeing any data.

At the end of each trial, participants made a forced-choice judgment selecting which of the five causal structures

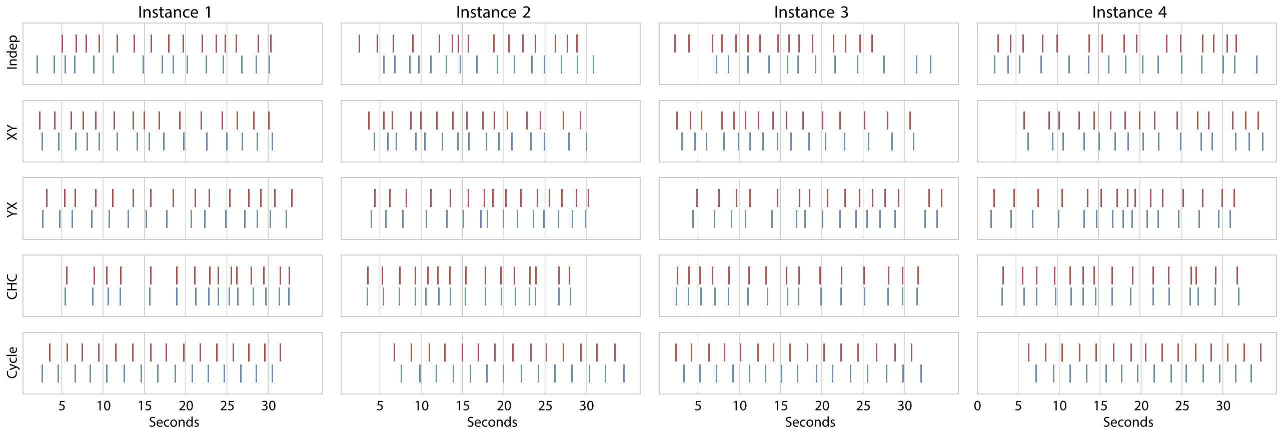


Fig. 4 All stimulus sequences presented to participants, where X events are denoted by red lines and Y events are denoted by blue lines

generated the data and provided a confidence in this judgment using a slider ranging from a left pole of 0 for "completely uncertain" and a right pole of 100 "completely certain" (with increments of 1, and without a starting value to minimize anchoring effects), as shown in Fig. 3.

The left/right position of the X and Y objects was randomized for each video (with the labels remaining in place), and the colors of the lights were randomly drawn for each trial to make different trials more distinguishable and reinforce the idea that different causal structures could govern different pairs of bacteria. Colors were sampled such that each pair had maximally dissimilar hues in a hue/saturation/value color space, but equal saturation and value (i.e., brightness). Events took the form of bacteria lighting up: This was displayed by having the colored circle representing the bacterium become maximally visible (by assigning minimal opacity) and then decay exponentially back to its baseline opacity, fading into the gray background with a rate of 25% per video frame (with a refresh rate of 50ms), providing a visual presentation that is consistent with the semantics of point events but ensures participants could easily perceive and distinguish between the events.

Stimuli

All video stimuli were generated by sampling from one of the five causal structures under consideration. Sampled delays were generated from distributions that provided adequate evidence for the true generative structure, while leaving some uncertainty. In particular, we set the parameters such that recurrence of each cause occurred with a longer, more variable delay than those delays between causes and effects, ensuring that there would be little chance of a cause recurring while its effects were still underway or of effect events overtaking each other. Specifically, for the independent structure, we set $\mu_a = 2.0s, \sigma_a^2 = 0.4s$ and $\mu_b = 2.0s, \sigma_b^2 = 0.4s$. For $X \rightarrow Y$, we set $\mu_a = 2.0s, \sigma_a^2 = 0.4s$ and $\mu_b = 0.5s, \sigma_b^2 = 0.01s$ (for $X \leftarrow Y$ swapping a and b). For the common hidden cause structure, we set $\mu_a = 2.0s, \sigma_a^2 = 0.4s$ and $\mu_b = 0.5s, \sigma_b^2 = 0.01s$. For the causal cycle, we set $\mu_a = 1.0s, \sigma_a^2 = 0.01s$ and $\mu_b = 1.0s, \sigma_b^2 = 0.01s$, which leads the self-delays of the X's and Y 's to be the same as for all other structures in expectation, while providing reliable cues. All stimulus parameterization settings are also reported in Appendix Table 2 as an overview. For each causal structure, we generated four stimulus sequences using the same generative model and selected one uniformly at random for each participant to average over idiosyncrasies of particular sampled sequences. All stimuli are visualized in Fig. 4 and the exact timings of the events are included in the pre-registration materials (see https://osf.io/e5d23).

Results

Overall, participants recovered the correct structure 58% of the time (compared to a chance-level accuracy of 20%). However, accuracy depended on the ground-truth causal structure. Figure 5 displays confusion matrices showing how often participants recovered the true generative structure in each condition. For the independent and the directed causal structures ($X \rightarrow Y$ and $Y \rightarrow X$), participants most frequently identified the ground truth. For the common hidden cause (CHC), people more often judged the data to be generated by a causal cycle, and for the causal cycle, people more often tended to favor one of the directed structures.

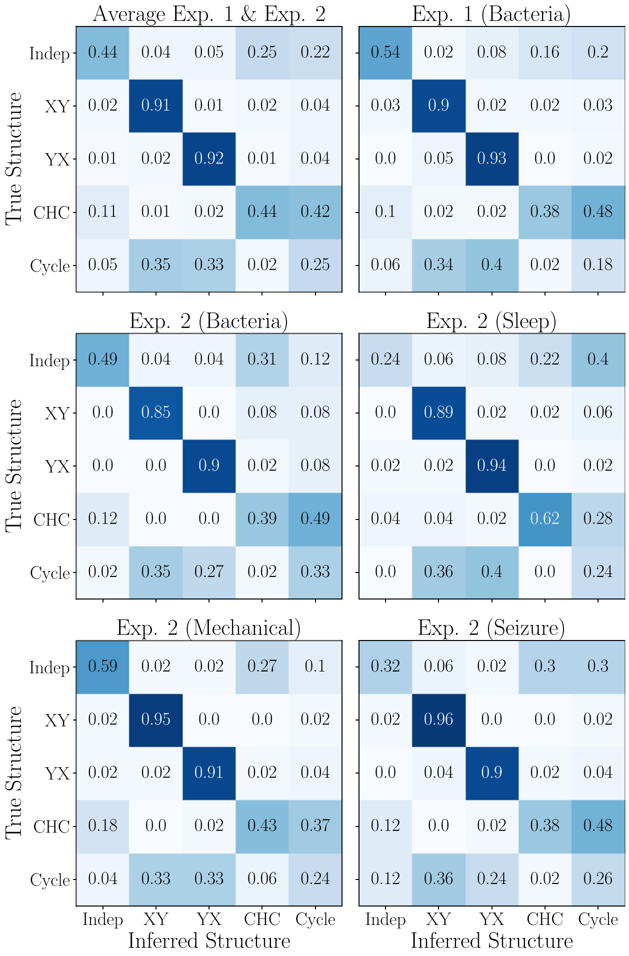


Fig. 5 Confusion matrices, containing people's judgment frequencies relative to ground truth for the respective experimental condition

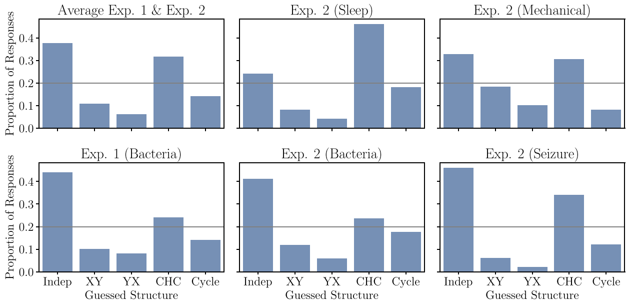


Fig. 6 Prior expectations (guess frequencies on trials without evidence) across all experiments ($N = 250$). Horizontal gray lines correspond to chance-level responding

Results for the prior elicitation condition are presented in Fig. 6. As indicated by deviations from the gray horizontal line, participants descriptively favored independent causal relationships and the common hidden cause over other possibilities.

Experiment 1 had a cover story of bioluminescent bacteria in a Petri dish, and it may be that the findings reported above express people's idiosyncratic expectations about the behavior of such bacteria. To assess the domain-generality of people's causal inferences from temporal data, we repeat (and internally replicate) the task under a range of different cover stories, as we report below.

Experiment 2

The aim of our second experiment was to replicate the findings from Experiment 1, and to understand (1) whether Experiment 1's results reflect domain-general expectations or specific expectations about our bacterium cover story; and (2) in the event that participants' beliefs vary, how they do and to what extent their elicited priors are consistent with their inferences. We included a replication of the original cover story and three new between-participant conditions with different cover stories. The experiment was pre-registered on OSF (https://osf.io/jq9bd).

Methods

Participants

Two-hundred adults recruited with Prolific Academic (113 female, 1 other; mean age 33.49 years, $SD = 11.59$) took part in Experiment 2 in return for a basic payment of £1.30 and performance-related bonuses of up to £1.20 resulting in a an overall average compensation of £10.49/h. Participants took 13.71 min on average ($SD = 6.30$) to complete the task. The experiment was conducted online with participants recruited via Prolific Academic (www.prolific.co). The criteria used to select participants were the same as for Experiment 1. The power calculation for determining the required sample size follows the rationale for Experiment 1, as 50 participants were allocated to each cover story condition.

Design and Procedure

Participants were randomly assigned to one of the four cover story conditions. These cover stories were selected to cover a wide spectrum of everyday inference domains and potentially different inductive biases, based on prior work that focused on capturing people's beliefs across a range of different domains (Yeung & Griffiths, 2015; Kushnir et al., 2010). In particular, the stories were selected to fulfill three desiderata: First, events and specifically their onsets can happen very quickly. Second, the causal "type" of the observed variables is the same (as opposed to, e.g., different medical symptoms following an infection). Finally, there are plausible mechanisms for all five causal structures in the setting described by the cover story.

Biological Bioluminescent bacteria, where events are illuminations of either bacterium. Identical to Experiment 1, serving as an internal replication.

Medical Recordings of mirco-seizures in animals' brains, where events correspond to local seizures on either side of the brain.

Mechanical Mechanical devices, where an event corresponds to one of the two knobs being pushed out (following Kushnir et al., 2010).

Behavioral Sleep recordings of two people, where events correspond to brief waking-up events of either person.

In all other ways, Experiment 2 was identical to Experiment 1.

Stimuli

Stimulus selection and sampling were identical to Experiment 1.

Results

Experiment 2 replicated the qualitative pattern observed in Experiment 1. As shown in Fig. 5, participants in the biological cover story again identified independent and directed causal structures correctly but tended to mistake the hidden cause for a cycle and the cycle for a unidirectional structure. However, there was some variation on this pattern in the three new cover stories, as we discuss below.

Results for the prior elicitation condition are presented in Fig. 6. We observe that the replication condition matches the qualitative pattern observed in Experiment 1. Meanwhile, there are qualitative differences in response patterns across cover story conditions, with participants descriptively reporting independent causes and a common hidden cause most often.

Inferential Statistics

Before turning to model-based analyses, we provide frequentist inferential statistics for both experiments. Overall, participants performed above chance (i.e., 20%), aggregating over the five causal structures (exact binomial test, $p < 0.001$). Running separate binomial tests per ground-truth structure with a Bonferroni-adjusted significance level of $\alpha = \frac{0.05}{5} = 0.01$ indicates better than chance-level performance for all conditions (independent $p < 0.001$, $X \rightarrow Yp < 0.001$, $X \leftarrow Yp < 0.001$, $X \leftarrow H \rightarrow Yp < 0.001$) except for when the true structure was a cycle ($p = 0.048$).

In order to test whether people's judgments differed between cover story conditions and ground-truth structures, we fit a multinomial logistic regression with the inferred causal structure as our criterion and the ground-truth causal structure as well as cover story as dummy- (or one-hot)-coded predictors. Likelihood-ratio tests for the two predictors reveal significant main effects of both ground-truth causal structure ($\chi^2(16) = 1657$; $p < 0.001$) and cover story ($\chi^2(12) = 29.913$; $p = 0.003$). Qualitatively, the main distinctive pattern here is that people in the sleep condition had a propensity to infer the CHC structure when the CHC was the ground truth, relative to other conditions. We conjecture that this effect is the result of there being a more salient and intuitive common hidden cause—noises waking both sleepers—in this condition.

Focusing on the prior elicitation trials over our cover stories in aggregate, we see that judgments deviated systematically from a random guesser null-hypothesis over the five candidate causal structures ($\chi^2(4) = 95.12$; $p < 0.001$). Meanwhile, the null-hypothesis that prior judgments are independent of the cover story could not be rejected ($\chi^2(12) = 19.168$; $p = 0.085$).

Model-Based Analyses

Having demonstrated that judgments are sensitive to and broadly aligned with the ground-truth causal structure, we now turn to our formal modeling framework to better understand whether participants' inferences are consistent with one or other of the Bayesian accounts we have outlined. That is, we will use model comparison to estimate to what extent participants' judgments are sensitive to delay information or just event order.

This requires (1) defining a prior over structures and priors over causal delay distributions; (2) computing marginal structure likelihoods and combining these with the structure prior; and (3) converting the resulting posterior into choice probabilities. We lay out each of these steps below.

Priors

Rational models of inference under uncertainty must generally ascribe a set of prior beliefs or inductive biases to the agents being studied. In a classical rational analysis, these priors are assumed to be matched to the statistics of environment to which agents are adapted (Anderson, 1991), but, as is the case in many models of causal learning, it is unclear how one would estimate these statistics directly. An alternative is to estimate participants priors empirically, and see whether participants' inferences from data are consistent with these priors. We compare such a prior over causal structures (henceforth the *elicited* priors), to a uniform *baseline*, as well as a *flexible* prior that is fitted to participants' judgments.

For the flexible prior, we fit a categorical prior distribution over our five causal structures along with the rest of our model. This required 4 parameters with the fifth element determined by the requirement that the prior sums to 1. For the elicited prior, we compute the relative choice frequencies of the causal structures in stimulus-free prior-elicitation trials across all participants. We assume guesses in the elicited prior condition represent samples obtained via probability-matching from the prior and that learners roughly share this prior; the choice frequencies can be taken as a reflection of this prior. In order to allow for some deviations from probability-matching choice behavior, either towards more random responding or harder maximization, we pass the elicited (log) prior through a softmax with a weight parameter $\omega \in [0, \infty]$ which we also fit to the training folds in our cross-validation procedure (as explained below).

Calculating Marginal Likelihoods—Including Conservatism

We assumed uninformative or weakly informative priors for all delay model parameters. For details on approximate inference for the delay model, see Appendix 4. We confirmed the stability of our marginal likelihood estimates by running the full procedure 5 times, which resulted in the same findings (yielding the same results at the numerical precision we report).

The order and delay models we introduced assign structure likelihoods for any sequence of X and Y events, for the delay model after marginalizing over potential hidden causes and parameters. However, these model predictions follow from the, arguably unrealistic, assumption that participants have perfect and arbitrarily precise perception and memory of temporal order and delays. We thus accommodate the possibility that participants have more uncertainty in their likelihoods than that assumed by our ideal-observer models. This is achieved by passing our order and delay model likelihoods through a softmax function with a fitted temperature λ. This allows for varying levels of perceptual uncertainty, possible memory failures, and conservatism with regard to belief change, conceptually following prior work (e.g., Edwards, 1968; Bramley et al., 2014). Intuitively, as λ becomes small, $\lambda \to 0$, we obtain stronger likelihoods, for $\lambda \to \infty$ we approach uniform (non-diagnostic) likelihoods, while the use of log probabilities allows us to also recover the original likelihoods.

Mapping Posterior Probabilities to Decisions

Participants do not provide posterior distributions directly but make forced-choice judgments about which of the five structure diagrams produced each observed sequence. Our decision model thus corresponds to a multinomial distribution over the five causal structures, where the probability of selecting each structure is given by the respective posterior structure probability. That is, the (normalized) product of structure prior and structure likelihood, as given in Eq. 1.

Cross-validation

We assess the predictive performance of order and the delay model variants on the judgments of all 250 participants. Excluding the prior elicitation trial, each participant made five judgments based on event sequences meaning there are 1250 choices, each made from the same set of five options. We evaluate the models' ability to predict participants' judgments for each of the ground-truth structures using softmax temperature parameters fit to data from the other four ground-truth structures. For the flexible prior models, we fit the four free prior parameters as well as the likelihood softmax temperature. For the elicited prior models, we fit a prior softmax temperature and the likelihood softmax temperature. In each case, we evaluate model performance in terms of the average cross-validated log-likelihood for the left-out ground-truth structure. This provides a rigorous test of generalization, as predictions are made towards an unseen experimental condition while avoiding potential issues with commonly used information criteria for model selection (e.g., Arlot & Celisse, 2010; Vehtari et al., 2017). As we have four different cover story conditions and do not posit systematic relationships between domains, we perform cross-validation on each cover story individually and subsequently aggregate the results. We also assess to what extent computing the elicited prior for each cover story separately or in aggregate aids in predicting peoples' judgments.

Model Comparison Results and Discussion

Table 1 presents aggregate (summed) cross-validation negative log-likelihood values for the different models. Overall, the delay likelihood with a uniform prior over causal structures gave the best predictions (as quantified by the lowest negative log-likelihood). The delay likelihood with a uniform prior outperformed all other models, including

Table 1 Model comparison results

Model name	Prior $p(s)$	Likelihood $p(\mathcal{D} \mid s)$	CV NLL
Baselines			
Random	Uniform	–	2011.80
Flexible	Flexible (4 free parameters)	–	2633.75
Elicited prior	Mean elicited prior	–	2020.19
Elicited prior per domain	Mean elicited prior per domain	–	2021.27
Alternative models			
Order	Uniform	Order	1704.22
Delay	Uniform	Delay	**1670.29**
Flexible order model	Flexible	Order	1980.64
Flexible delay model	Flexible	Delay	2187.49
Elicited prior order	Avg. elicited prior	Order	1704.22
Elicited prior delay	Avg. elicited prior	Delay	**1670.29**
Elicited prior per domain order	Avg. elicited prior per domain	Order	1712.47
Elicited prior per condition delay	Avg. elicited prior per condition	Delay	1672.82

Model comparison using cross-validated negative log-likelihoods (NLL) of the respective decision model on all experimental data, where model predictions are made towards judgments from unseen ground-truth structure conditions. Best predictive performance (lowest NLL) in boldface

all baselines and other combinations of priors with order or delay likelihoods. In particular, incorporating elicited priors did not lead to better cross-validated predictions, neither when using the average elicited prior across cover stories nor with elicited priors computed separately by cover story. We note that the delay likelihood with elicited priors resulted in the same minimal negative log-likelihood, which can be explained by a very large value for the prior softmax temperature, which essentially transforms the elicited prior into a uniform prior. These results are in line with the idea that people use fine-grained delay information to infer causal structure rather than relying just on the order the observed events occurred in. Figure 7 presents posterior probabilities per ground-truth causal structure when fitting the order and delay likelihood softmax temperature on all data simultaneously. The fitted values are $\lambda = 4.02$ and $\lambda = 3.94$ for order and delay model predictions, respectively. We discuss the finding that elicited priors did not aid in predictions in the general discussion.

Regarding model recovery, for the event sequences studied here, the delay model always recovered ground-truth structure in our simulations (across five random runs), with the softmax transformation providing a slightly softened version of the original model predictions. The order model deviated systematically, as can also be seen in Fig. 7. This structure mis-identification of the order model is not surprising, as the Bayesian Ockham's razor penalizes structures whose predictions are less specific to the observed data. Specifically, a cyclic (alternating) sequence is also consistent with a unidirectional structure, where the directionality is determined by which variable activates first, but the latter is less flexible. The cyclic structure spreads its likelihood across sequences beginning with X or Y, leading to unidirectional chains to be inferred instead under an order-only account. We observe a similar trend towards misidentifying causal cycles in the human judgments and elaborate on this further in the general discussion.

Individual Differences

Even though our aggregate results indicate that, overall, participants were best accounted for by the delay model, we may expect that people differ in their use of delay statistics over order heuristics, as well as how reliably they choose structures that are best supported by the evidence, as captured by

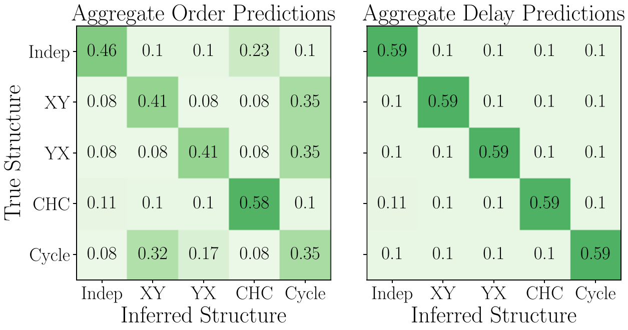


Fig. 7 Posterior probabilities of the computational models for the stimuli sequences displayed to participants when fitting aggregate data, comprising all judgments from Experiment 1 and Experiment 2

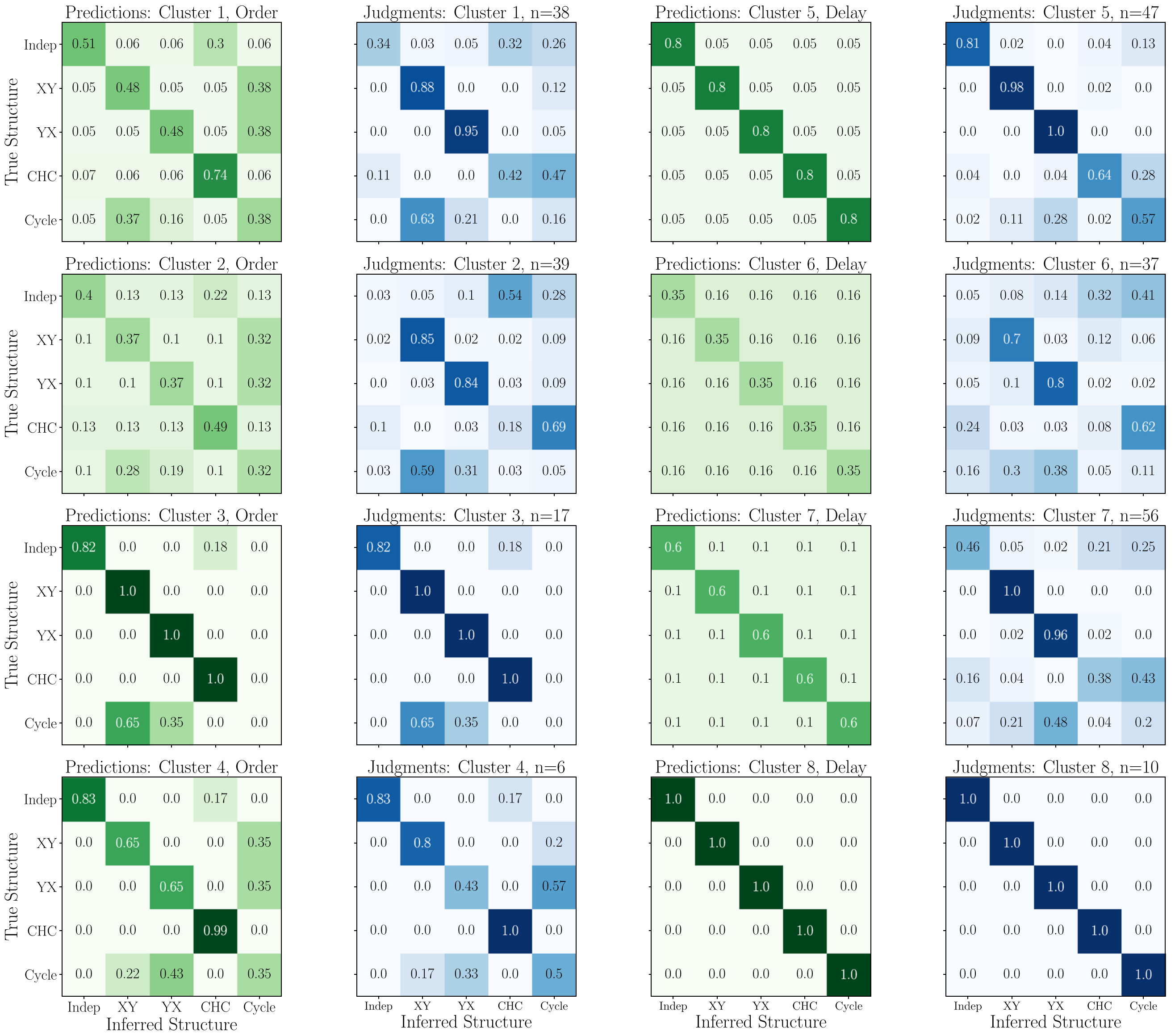


Fig. 8 Posterior model predictions and confusion matrices of people's judgments for all clusters

the softmax temperature described above. To better understand those individual differences, we used a probabilistic mixture model to identify different clusters of participants, and see which models (delay, order), and parameters (temperature) best explain the behaviors in each cluster.

Specifically, we computed (hard) cluster assignments via expectation maximization (Dempster et al., 1977) conditional on numbers and types of clusters, and chose the number of order and delay clusters using the Bayesian information criterion (BIC) to penalize the complexity and flexibility inherent in larger numbers of clusters. The number of clusters was allowed to vary between 0 and 7 for order and delay clusters each, resulting in an overall number of clusters between 1 and 14. We re-ran the procedure 30 times with different initial parameters to mitigate getting stuck in local minima.

We find the lowest BIC for four order and four delay clusters (BIC = 2693.76). We present posterior model predictions for each cluster in Fig. 8 as well confusion matrices of human judgments in each cluster. As presented in Appendix Fig. 16, the minimum is relatively flat, so that we do not make strong claims about the exact number of clusters to describe a general population. Looking at judgments per cluster in Fig. 8, we can find several qualitative patterns: First, many participants are best described by a strictly or nearly normative delay account (47 and 10 participants in clusters 5 and 8, respectively). Overall, 150 out of 250 participants are better described by delay model predictions than an order-only account. Second, several participants tended to infer a unidirectional relationship when the ground truth was a cycle, typically treating the first event to occur in the sequence as the cause. These participants are allocated to both order and delay clusters (in particular clusters 1, 2, 3, 6, and 7). We discuss this effect in the "First-Mover Effects" section. A third pattern is that several participants inferred cycles when the ground truth was given by a common hidden cause (clusters 1, 2, 6, and 7), also discussed in the "First-Mover Effects" section. Lastly, there are a number of participants who infer common hidden causes when events are independent (clusters 2 and 6, in particular). This may be attributable to a simplicity preference, in which people preferentially posit fewer explanatory variables (Lombrozo, 2007).

Overall, this analysis indicates that a large proportion of people make inferences that are well-aligned with our normative expectations, and exploit both delay and order information to identify causal structure and hidden causes. At the same time, we have observed phenomena that raise the question of how one may enrich models of causal reasoning with temporal dynamics to capture the breadth of inferences people make; we discuss this further below.

General Discussion

The present results expand our understanding of how people exploit temporal information to identify causal structure. Our findings demonstrate that people use temporal cues including both order and delays to make generally appropriate inferences about structures that include hidden causes and cyclic relationships. We found that people make inferences that are broadly consistent with Bayesian inference from an uninformative prior based on inter-event intervals, doing so similarly across four domains. Furthermore, we considered a range of different assumptions that might drive these inductive inferences. We found that people's structure judgments show a sensitivity not just to the order in which events occur, but to the distribution of inter-event delays, using a comparison of several probabilistic models, extending results from previous research (Bramley et al., 2018). We did not find that people's elicited priors help in predicting their judgments. Furthermore, there were some notable differences between model predictions and participants' judgments, with implications about (1) ways people might make systematic errors in real-world causal inference problems; and (2) different assumptions people make about causal structures with temporal dynamics that are not captured by current accounts.

First-Mover Effects

In almost all combinations of cover story and structure, participants inferred the true generative structure more often than any alternative, with one conspicuous exception: participants more often judged causal cycles to be unidirectional causal links (X causes Y or Y causes X). What could explain this finding? One key consideration may be whether participants interpret the beginning of the video clips as an observed "window" of an ongoing stream of events or as the actual beginning of the causal system dynamics. One possibility is the existence of a "first-mover-effect," i.e., that people view the first event they notice in an alternating sequence (with symmetric delays) as the "root-cause" of the series of events, rather than inferring a cycle. We tested whether when participants mis-identified a causal cycle, they were most likely to mistake it for a directed model in which the cause is the first variable to activate. This turned out to be the case 96% of the time, a significant deviation from chance (χ^2 test of independence indicates $\chi^2(1) = 142.235$; $p < 0.001$). Meanwhile, this inference error would also be partly consistent with an order-only account, as the likelihood for the cycle is spread across either variable starting the sequence. From this perspective, people may interpret the alternating sequence as a repeated activation of the cause on its own (i.e., the variable starting the observed sequence), which triggers the effect events. Alternatively, people may have inferred that there is a cycle, but simply reported on which variable activated first, thereby starting the cycle.[4]

Future work may probe these potential biases further, for instance assessing real-world situations in which a participant is observing an ongoing symmetric cyclic process and views one component as the cause, simply because it occurred when one started observing the process. It would further be useful to investigate to what extent this finding can be explained by delay accounts with a first-mover bias or an order-only account, for instance by altering the stimulus delay parameters or providing a "fade-in" period, where the start of the video clips is slowly increasing in luminescence, or otherwise making it clear that learners are observing a causal system that is already "running" when participants start observing it. Additionally, it could be tested whether the effect may be more pronounced when only little data is available and diminish with longer stimulus sequences. This departure from model predictions warrants further investigation and raises the question of inductive bias, that is, to what extent human judgments depend on background knowledge and domain-specific expectations about the causal system under consideration.

[4] It is conceivable that participants have a preference for a lexicographic order with X being the cause and Y the effect. However, a test for whether there is a difference in the marginal frequency of how often $X \rightarrow Y$ or $Y \rightarrow X$ are inferred revealed no such effect (two-tailed exact binomial test, $p = 0.624$), which aligns with a lack of qualitative trends, as shown in Appendix 9.

Hidden Causes or Causal Cycles

Another pattern that emerged in some clusters was a tendency to choose causal cycles when the ground truth was a common hidden cause. One explanation for this observation is that some participants, rather than treating the events as a single continuous stream, may partition events—especially when these events occur in bursts separated by longer delays—as sets of distinct episodes. In this scenario, there are cases where Y follows X after a very short delay—suggesting an $X \rightarrow Y$ edge—as well as cases where X follows Y after a very short delay—suggesting a $Y \rightarrow X$ edge. In combination, these edges imply a causal cycle. Exploring this possibility is an interesting prospect for the future, but doing so would require an account of how and when people partition continuous time sequences into discrete events, and potentially more complex causal models.

Seeing Structure in Independent Event Sequences

Participants sometimes inferred a relationship between X and Y when the two variables were independent, particularly viewing CHC and cycles as plausible explanations in the independent conditions. What explains this finding of people seeing structure where there is none? One possibility is that people might infer a CHC when there are several instances of X and Y occurring in close temporal proximity, perhaps to the point of perceiving simultaneous events. Alternatively, if events are farther apart in time, people may infer a causal cycle, following the speculation above that people partition sets of events into episodes. Additionally, a prior belief about a CHC with a an unreliable delay distribution from hidden to observed events could also lead to patterns that resemble independent sequences. We did not observe distinctive differences in inferences for the different instances of independent sequences presented to participants, and a systematic investigation of this mis-identification would require more instances of stimulus sequences, possibly with explicit manipulations. From the modeling side, exploring this further could be aided by relaxing the assumptions that causes invariably lead to their effects, and that there are no outside influence beyond those presented in the causal structures, as we also discuss below. More broadly, previous work has established that people sometimes see structure where there is none (Williams & Griffiths, 2013; Griffiths & Tenenbaum, 2007; Blanco, 2017), and explanations from this line of research may be fruitful directions for future theoretical work.

Hidden Causes with Tied Parameters

Our delay model assumed that the delay distributions mapping from hidden causes to observed variables have tied parameters, encoding the belief that they share a common causal mechanism. For example, if two bioluminescent bacteria are responding to the same environmental stimulus, e.g., a shock, is it plausible that the mechanisms driving their responses are similar. Consequently, we might expect that the distribution of shock-to-response delays across these bacteria resembles (or matches) the distribution of delays for a single bacterium responding to multiple shock events over time; this is naturally formalized by coupling the parameters of the delay distributions. In contrast, if we have a bacterium and a mechanical sensor that are responding to the same shock event, there is little reason to infer a similar mechanism is at work and it may thus be more reasonable to assume distinct, uncoupled distributions for the two delay distributions. Between these two extremes, one might assume a weak coupling between the parameters, e.g., implemented via a hierarchical model (Lucas & Griffiths, 2010).

While the CHC with tied parameters leads to an expectation of observed events being clustered together in time (with no systematic temporal precedence of one event type over another), allowing for untied delay distributions creates identifiability issues. For instance, a CHC may look like a unidirectional structure between the observations when the delay distributions differ in their expectation and have sufficiently low variance to follow an alternating order. Generally, untied parameters mean that the CHC structure receives an additional set of parameters and this increased complexity is penalized due to the Bayesian Ockham's razor (Myung & Pitt, 1997).

As a supplementary analysis, presented in Appendix 7, we assessed a delay model that did not assume tied parameters for the CHC structure. As expected, untied parameters lead to lower posterior probability for the CHC when the CHC is the ground truth, but do not change the results for the other structures at the numerical precision we report. Importantly, this mis-identification does not seem to capture people's deviations from our predictions, and leads to lower scores in the global cross-validated model comparison as compared to our original results with tied CHC parameters. However, exploring this assumption with manipulated context information may be a valuable direction for future work, in particular in relation to prior expectations about weak coupling between parameters, as could be formalized using a hierarchical model.

Representing Inductive Biases

A general question in human causal learning concerns the role of domain-specific expectations, e.g., about what structures are plausible in particular domains, and finer details such as how quickly or reliably a cause might bring about its effect. In our experiments, neither elicited priors nor fitted priors over causal structures helped in predicting people's judgments when combined with delay model likelihoods (Table 1). This is consistent with participants having weak and idiosyncratic beliefs about plausible causal structures, but in light of our sample size and the observation that there were systematic preferences in our prior elicitation condition, it was nonetheless surprising that elicited priors did not help in predicting people's judgments.

How can we explain this mismatch, which, at first glance, could be taken to suggest that—contra many previous accounts of causal reasoning—people do not integrate their prior beliefs and evidence in a probabilistically coherent way. One possibility is that our assumption that people would probability-match in the elicitation condition, i.e., make discrete choices with probability proportion to their degrees of subjective belief (e.g., Costello & Watts, 2014; Acerbi et al., 2014), was unwarranted—perhaps participants' judgments reflected a policy more like maximization, and their real prior beliefs were nearly uniform. An alternative possibility is that our within-participant design led to systematic effects that were orthogonal to participants actual prior beliefs: That is, contra our instructions but consistent with our actual design (and common practice), participant may have expected to see each causal structure at least once. Additionally, the influence of context information on people's inference may come to bear only when people have stronger mechanistic intuition about the causal relationships and the cover stories we considered may not be associated with sufficiently strong prior beliefs. Future work, including between-subjects designs for prior elicitation and manipulations drawing on the literature on prior elicitation, also in statistics (e.g., Stefan et al., 2022; Barrera-Causil et al., 2019), and work on the role of context information in causal cognition (Buehner & May, 2002; Buehner & McGregor, 2006b; Buehner & May, 2003; Griffiths & Tenenbaum, 2009; Schlottmann, 1999; Schlottmann et al., 2002) may help resolve this issue and shed light on causal cognition more generally. While our analysis and the previous discussion largely focused on structure-level biases, people might plausibly expect different causal mechanisms—and, by extension, cover stories—to entail different delay distributions. Exploring people's inductive biases about delay distributions in causal mechanisms and how they inform structural inferences is an important direction for future work.

Towards Richer Causal Models

For the sake of simplicity, all of our experiments dealt with event types that had at most one cause, e.g., a bacterium might illuminate due to a hidden cause or the activity of the other bacterium, but not both. Additionally, in the present study, our generative model was stochastic in terms of when causes brought about their effects, but influences were deterministic in their causal strengths. That is, the occurrence of a cause invariably led to the occurrence of its effect. In the real world, however, causes can sometimes fail, thereby introducing additional complexity and ambiguity into the learning task. We also make the assumption that there are no additional outside causes beyond the ones posited by the respective causal structure that can trigger effect events (except for the common hidden cause that invariably affects both X and Y). Studying human learning in the face of stochastic causal relationships and variable delays would thus be an interesting direction for future research, but would also severely complicate the inference problem, which is already challenging, as we are marginalizing over all unknown quantities.

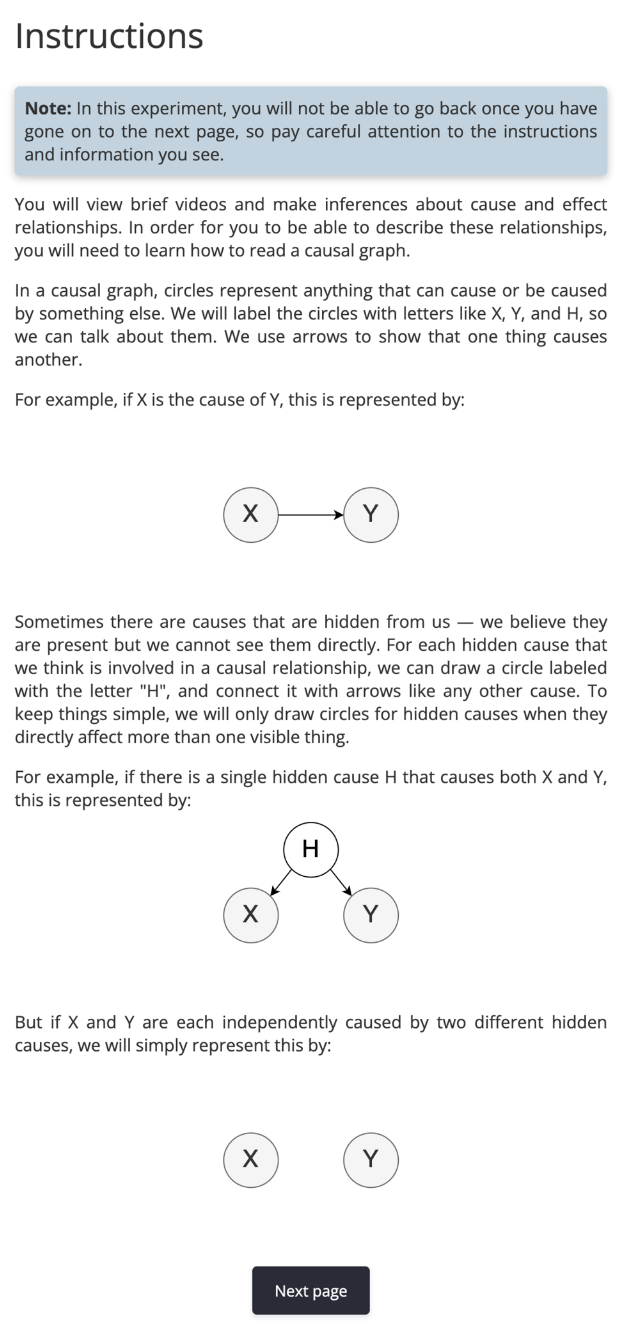
Furthermore, we dealt with a strictly observational setting and future work may probe how people intervene on dynamic causal systems (e.g., Gong et al., 2022) when the set of causal structures may include hidden causes, following a rich literature on interventions in static human causal learning (Steyvers et al., 2003; Coenen et al., 2015; Bramley et al., 2015).

More generally, future work may benefit from including conditions with alternative judgment elicitation paradigms, such as free-form drawing of nodes and edges to describe causal relationships, as well as qualitative, natural language, descriptions.

Conclusions

We have presented the first study investigating the role of temporal information in how people discover hidden causes and causal cycles. We conducted two novel experiments that covered different domains and tested several computational models. We took an ideal-observer perspective to examine how people go from observed sequences of events to inferred structures. We found participants were broadly consistent with a Bayesian account, and were able to use order and timing information to identify causal structure, including the presence of a common hidden cause where appropriate. Meanwhile, several groups of participants showed systematic patterns of judgments that deviated from all of our models, which suggests future experiments as well as opportunities for enriching the causal models we attribute to human learners. These findings expand our understanding of how people learn about causal structure from an ongoing stream of observed events with temporal dynamics and open up several potentially fruitful avenues for future research.

Appendix: 1. Participant training and comprehension checks



Instructions

Note: In this experiment, you will not be able to go back once you have gone on to the next page, so pay careful attention to the instructions and information you see.

You will view brief videos and make inferences about cause and effect relationships. In order for you to be able to describe these relationships, you will need to learn how to read a causal graph.

In a causal graph, circles represent anything that can cause or be caused by something else. We will label the circles with letters like X, Y, and H, so we can talk about them. We use arrows to show that one thing causes another.

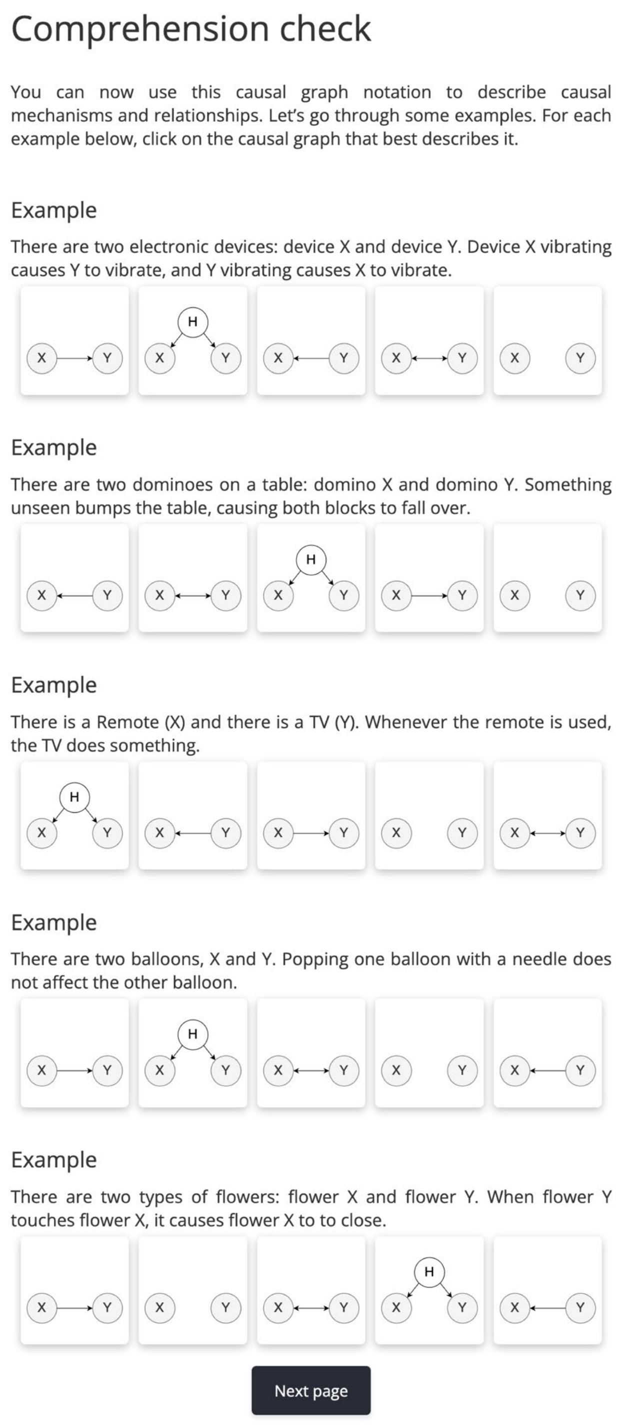
For example, if X is the cause of Y, this is represented by:

Sometimes there are causes that are hidden from us — we believe they are present but we cannot see them directly. For each hidden cause that we think is involved in a causal relationship, we can draw a circle labeled with the letter "H", and connect it with arrows like any other cause. To keep things simple, we will only draw circles for hidden causes when they directly affect more than one visible thing.

For example, if there is a single hidden cause H that causes both X and Y, this is represented by:

But if X and Y are each independently caused by two different hidden causes, we will simply represent this by:

Fig. 9 Screenshot of page of online experiment that explained causal graphical models



Comprehension check

You can now use this causal graph notation to describe causal mechanisms and relationships. Let's go through some examples. For each example below, click on the causal graph that best describes it.

Example

There are two electronic devices: device X and device Y. Device X vibrating causes Y to vibrate, and Y vibrating causes X to vibrate.

Example

There are two dominoes on a table: domino X and domino Y. Something unseen bumps the table, causing both blocks to fall over.

Example

There is a Remote (X) and there is a TV (Y). Whenever the remote is used, the TV does something.

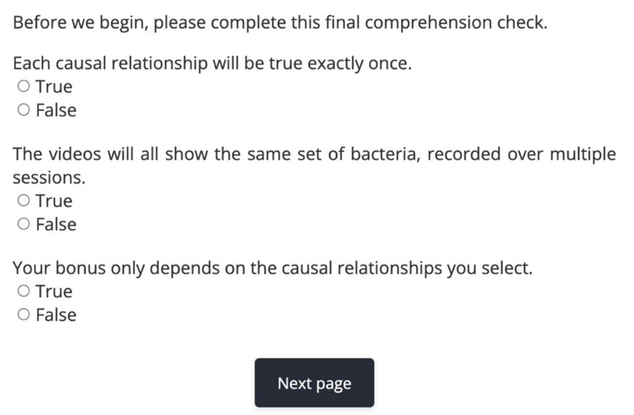
Example

There are two balloons, X and Y. Popping one balloon with a needle does not affect the other balloon.

Example

There are two types of flowers: flower X and flower Y. When flower Y touches flower X, it causes flower X to to close.

Fig. 10 Comprehension check 1, which participants had to complete successfully before progressing in the experiment



Before we begin, please complete this final comprehension check.

Each causal relationship will be true exactly once.
○ True
○ False

The videos will all show the same set of bacteria, recorded over multiple sessions.
○ True
○ False

Your bonus only depends on the causal relationships you select.
○ True
○ False

Fig. 11 Comprehension check 2, which participants had to complete successfully before progressing in the experiment

Appendix: 2. Cover story styling

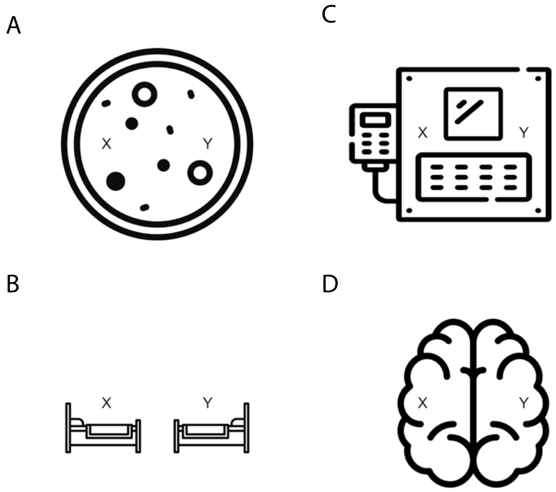


Fig. 12 Backgrounds used for stimulus displays for different cover stories. (*A*) Biological (bacteria); (*B*) behavioral (sleep); (*C*) mechanical (devices); (*D*) medical (micro-seizures)

Appendix: 3. Order model representation

As time is continuous and events are treated as point events, the probability of two events occurring at exactly the same time is zero. However, as discretization is required to represent time on a computer as well as to display stimuli to participants, and taking into account that human perception is limited in its temporal precision, events that are very close in time might be perceived as simultaneous. In order to be consistent with the order-only account, we are thus required to marginalize over the two possible orderings behind observed simultaneous events for computing the structure likelihood. See Fig. 13 for graphical representations of the different causal structures under the order model as probabilistic finite state machines.

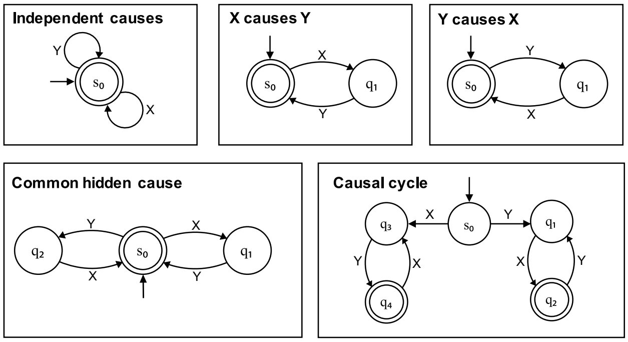


Fig. 13 Probabilistic finite state machines representing the different causal structures under and order-only account

Appendix: 4. Delay model

Delay model representation

For all delays, we assume uninformative or weakly informative priors, where the only fixed assumptions are that effects follow causes, and in the absence of information to the contrary, shorter delays are more likely than longer ones. Following Bramley et al. (2018), we here use a mean and shape parameterization of all gamma delays for expressing the prior distributions. We assume exponential (maximum-entropy) prior distributions with a scale of 10 for all mean and scale parameters. Similarly, for initial states, we assume exponential priors with a scale of 1. These priors capture the idea that people can learn that delay distributions are longer or shorter and more or less variable in a particular context. For computing $p(\mathcal{D} \mid s)$, we employ a Monte Carlo approach. As a simple Monte Carlo estimate of the marginal likelihood is unlikely to succeed with the parameter dimensionality, we approach this problem with importance sampling with parameterized proposal distributions. For all structures, we compute posteriors over delay parameters using Stan (Carpenter et al., 2017), with 4 chains, a burn-in period of 400 samples and 2000 MCMC samples. We then computed the mean and variance of the posterior samples of each parameter (that is, the marginal posterior mean and variance), which we use to parameterize independent gamma proposal distributions. We then use these parameterized gamma distributions as our proposal distributions in the importance sampling procedure for computing the marginal likelihood for a given structure (see, e.g., Murphy, 2012). For each sequence of events we use 10^6 importance samples. Note that in the case of the hidden cause structure, we further need to marginalize over the times at which the hidden events occurred. We marginalize over the hidden states in an inner Monte Carlo loop with an additional 10 samples to improve the stability of our estimates. See Fig. 14 for graphical representations of the different causal structures under the delay model.

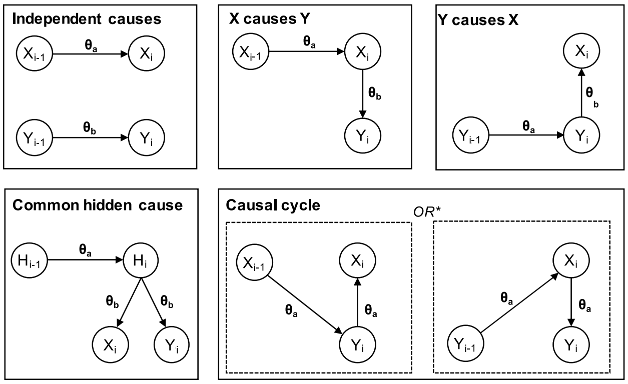


Fig. 14 Delay model as a DBN model. θ denotes parameters $[\mu,\sigma^2]^\top$; OR* indicates that the cycle can either start with X or Y, depending on how the initial state was sampled

Gamma distribution

Figure 15 shows density functions for different parameter settings of gamma distributions. More formally, the probability density function for the gamma distribution Gamma(k,θ) describing the delay x, where k is the shape and θ the scale parameter, is given by

$$p(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, \tag{4}$$

with Γ being the gamma function. We have that $\mathbb{E}[x] = k\theta$ and $\mathbb{V}[x] = k\theta^2$, and we can reparameterize the distribution with $\theta = \frac{\mathbb{V}[x]}{\mathbb{E}[x]}$ and $k = \frac{\mathbb{V}[x]}{\theta^2}$ for a more intuitive interpretation in the setting of causal delays. See Table 2 for the parameterizations used to generate the stimuli in our experiments.

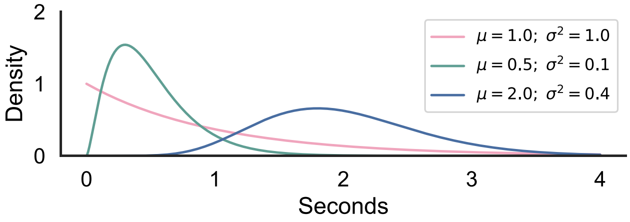


Fig. 15 Density functions of gamma distributions with different means and variances. The red line shows the special case in which the gamma distribution recovers one particular exponential distribution

Table 2 Parameterization for sampled event sequences used as stimuli for human experiments

Structure	μ_a	σ_a^2	μ_b	σ_a^2
Independent	2.0	0.4	2.0	0.4
$X \rightarrow Y$	2.0	0.4	0.5	0.01
$X \leftarrow Y$	0.5	0.01	2.0	0.4
CHC	2.0	0.4	0.5	0.01
$X \leftrightarrow Y$	1.0	0.01	1.0	0.01

For the mapping of labels a and b to causal delays, see Fig. 14

Appendix: 5. Confidence ratings

Table 3 Confidence ratings per ground-truth causal structure

			Confidence ratings						
True structure	Inferred correct	Count	Mean	SD	Min	First quartile	Medial	Last quartile	Max
Indep	No	141.00	55.86	23.32	0.00	40.00	53.00	74.00	100.00
	Yes	109.00	61.70	24.55	0.00	45.00	62.00	83.00	100.00
XY	No	23.00	60.09	23.12	6.00	46.50	57.00	75.50	100.00
	Yes	226.00	79.06	19.65	5.00	69.25	82.00	96.00	100.00
YX	No	21.00	62.76	21.00	21.00	50.00	65.00	77.00	100.00
	Yes	230.00	79.42	18.76	24.00	67.25	82.50	96.75	100.00
CHC	No	140.00	67.56	21.81	4.00	53.75	69.00	84.50	100.00
	Yes	110.00	59.90	22.93	0.00	46.25	58.50	77.75	100.00
Cycle	No	187.00	73.96	22.29	8.00	61.00	76.00	93.50	100.00
	Yes	63.00	62.46	23.72	5.00	47.50	65.00	81.00	100.00
Prior elicitation		250.00	12.92	23.17	0.00	0.000	0.00	16.00	100.00

Appendix: 6. Cluster analysis

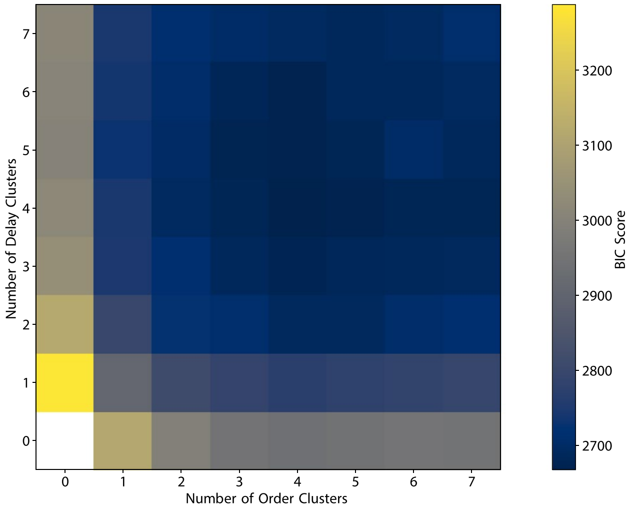


Fig. 16 BIC score as a function of the number of order and delay clusters, analysis comprising judgments from all 250 participants

Appendix: 7. Common hidden cause with untied parameters

As an additional analysis, we re-ran our evaluation procedure with a modified delay model with untied delay distribution parameters for the delays from the common hidden cause to the observables. Posterior probabilities for the respective causal structure on all stimuli from Experiments 1 and 2 are presented in Fig. 17 and Table 4. Overall, we observe worse model fits than for the delay model with tied parameters as reported in the main text.

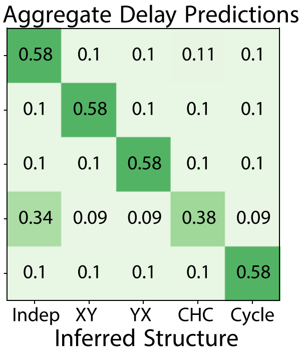


Fig. 17 Posterior probabilities for delay model with untied common hidden cause effect delays for the stimuli sequences displayed to participants when fitting aggregate data, comprising all judgments from Experiment 1 and Experiment 2

Table 4 Model comparison results for common hidden cause with untied parameters

Model name	Prior $p(s)$	Likelihood $p(\mathcal{D} \mid s)$	CV NLL
Baselines			
Random	Uniform	–	2011.80
Flexible	Flexible (4 free parameters)	–	2633.75
Elicited prior	Mean elicited prior	–	2020.19
Elicited prior per domain	Mean elicited prior per domain	–	2021.27
Alternative models			
Order	Uniform	Order	1704.22
Delay (untied)	Uniform	Delay	1718.59
Flexible order model	Flexible	Order	1980.64
Flexible delay model	Flexible	Delay	2278.80
Elicited prior order	Avg. elicited prior	Order	1704.22
Elicited prior delay (untied)	Avg. elicited prior	Delay	1718.59
Elicited prior per domain order (untied)	Avg. elicited prior per domain	Order	1712.47
Elicited prior per condition delay (untied)	Avg. elicited prior per condition	Delay	1720.39

Model comparison using cross-validated negative log-likelihoods (NLL) of the respective decision model on all experimental data, where model predictions are made towards judgments from unseen ground-truth structure conditions. Best predictive performance (lowest NLL) in boldface

Appendix: 8. Self-avoidance

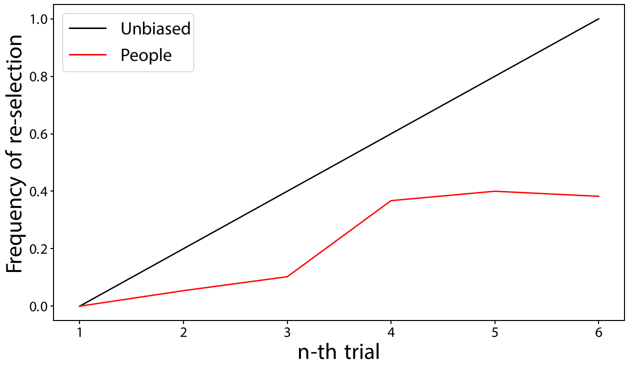


Fig. 18 Self-avoidance: relative frequency of structure re-selection on the respective trial number. Normative (unbiased) responding would have resulted in a pattern corresponding to the black line

Appendix: 9. Asymmetries between *X* → *Y* and *Y* → *X* inferences

Table 5 Frequencies the respective structures were inferred, depending on whether the left-right position was flipped or not

True structure	Inferred structure	Flipped	Not flipped
XY	XY	116	110
XY	YX	2	0
YX	XY	3	3
YX	YX	118	112

Acknowledgements We thank Tianwei Gong for feedback and valuable discussion about this project. We also thank Simon Stephan and another anonymous reviewer for constructive feedback on an earlier version of this manuscript.

Author Contribution SV, NRB, and CGL designed the study. SV carried out the experiments. SV analyzed the data and implemented the models. SV, NRB, and CGL wrote the manuscript.

Funding SV was supported by a Principal's Career Development Scholarship, awarded by the University of Edinburgh and received funding from the Institute for Language, Cognition and Computation at the University of Edinburgh. CGL and NRB are supported by an EPSRC grant (EP/T033967/1).

Data Availability Anonymized data generated in this study are hosted on the Open Science Framework https://doi.org/10.17605/OSF.IO/54VGF.

Code Availability Code for the analyses can be retrieved from: https://github.com/simonvalentin/discovering_common_causes.

Declarations

Ethics Approval Ethical approval was granted by the University of Edinburgh Informatics Research Ethics Panel, RT2019/58792.

Consent for Publication All participants included in the study provided electronic consent to their data being used in scientific publications.

Consent to Participate Informed consent to participate was obtained in electronic form from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

Authors' note: The work was completed prior to Simon Valentin joining Amazon.

References

Acerbi, L., Vijayakumar, S., & Wolpert, D.M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, *10*(6), e1003661.

Ahn, W.k., Kalish, C.W., Medin, D.L., & Gelman, S.A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352.

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Barrera-Causil, C.J., Correa, J.C., & Marmolejo-Ramos, F. (2019). Experimental investigation on the elicitation of subjective distributions. *Frontiers in Psychology*, *10*, 862.

Blanco, F. (2017). Positive and negative implications of the causal illusion. *Consciousness and Cognition*, *50*, 56–68.

Bramley, N., Gerstenberg, T., & Lagnado, D. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 36.

Bramley, N.R., Lagnado, D.A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Bramley, N.R., Dayan, P., Griffiths, T.L., & Lagnado, D.A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Bramley, N.R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D.A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12), 1880–1910.

Buehner, M.J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*(4), 269–295.

Buehner, M.J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, *56*(5), 865–890.

Buehner, M.J., & McGregor, S. (2006a). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking and Reasoning*, *12*(4), 353–378.

Buehner, M.J., & McGregor, S. (2006b). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*(4), 353–378.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20.

Coenen, A., Rehder, B., & Gureckis, T.M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Cook, C., Goodman, N.D., & Schulz, L.E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463.

Davis, Z.J., Bramley, N.R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244.

Dean, T., & Kanazawa, K. (1989). A model for reasoning about real-time processes. *Computational Intelligence*.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 245–264.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, *3*(2), 81–91.

Edwards, W. (1968). Conservatism in human information processing. Formal representation of human judgment.

Einhorn, H.J., & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Fernando, C. (2013). From blickets to synapses: Inferring temporal causal networks by observation. *Cognitive Science*, *37*(8), 1426–1470.

Friston, K.J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19* (4), 1273–1302.

Garcia, J., Ervin, F.R., & Koelling, R.A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, *5*(3), 121–122.

Gershman, S.J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N.R. (2022). Active causal structure learning in continuous time. PsyArXiv https://doi.org/10.31234/osf.io/jg2c5

Gopnik, A., & Tenenbaum, J.B. (2007). Bayesian networks, Bayesian learning and cognitive development.

Gopnik, A., Schulz, L., & Schulz, L.E. (2007). *Causal learning: Psychology, philosophy, and computation*. London: Oxford University Press.

Granger, C.W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Greville, W.J., & Buehner, M.J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.

Grice, G.R. (1948). The relation of secondary reinforcement to delayed reward in visual discrimination learning. *Journal of Experimental Psychology*, *38*(1), 1.

Griffiths, T.L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T.L., & Tenenbaum, J.B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Griffiths, T.L., & Tenenbaum, J.B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661.

Guo, R., Cheng, L., Li, J., Hahn, P.R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, *53*(4), 1–37.

Heinze-Deml, C., Maathuis, M.H., & Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, *5*, 371–391.

Hume, D. (1740). *A treatise of human nature*. Oxford Philosophical Texts (2000 reprint).

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT Press.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*(9), 678–683.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 25.

Kushnir, T., Gopnik, A., Lucas, C., & Schulz, L. (2010). Inferring hidden causal structure. *Cognitive Science*, *34*(1), 148–160.

Lagnado, D.A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Lagnado, D.A., & Sloman, S.A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(3), 451–460.

Lagnado, D.A., & Speekenbrink, M. (2014). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*(1), 184–195.

Lauritzen, S.L., & Richardson, T.S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(3), 321–348.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55* (3), 232–257.

Löwe, S., Madras, D., Zemel, R., & Welling, M. (2020). Amortized causal discovery: Learning to infer causal graphs from time-series data. arXiv: 200610833

Lucas, C.G., & Griffiths, T.L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Lucas, C.G., Holstein, K., & Kemp, C. (2014). Discovering hidden causes using statistical evidence. In *Proceedings of the annual meeting of the cognitive science society*.

Malinsky, D., & Spirtes, P. (2019). Learning the structure of a nonstationary vector autoregression. In *The 22nd international conference on artificial intelligence and statistics, PMLR* (pp. 2986–2994).

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. W.H. Freeman.

Mastakouri, A.A., Schölkopf, B., & Janzing, D. (2021). Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In *International conference on machine learning, PMLR* (pp. 7502–7511).

Michotte, A. (1946). La perception de la causalité. Louvain: Publications Universitaire.

Murphy, K.P. (2012). *Machine learning: A probabilistic perspective (adaptive computation and machine learning series)*. Cambridge: MIT Press.

Myung, I.J., & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Nodelman, U., Shelton, C.R., & Koller, D. (2002). Continuous time Bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

Pacer, M., & Griffiths, T. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In *Proceedings of the 37th annual conference of the cognitive science society (CogSci'15)*.

Pacer, M.D., Griffiths, T.L., & Griffiths, L. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th annual meeting of the cognitive science society (Cogsci2012)* (pp. 833–838).

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020). Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics, PMLR* (pp. 1595–1605).

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. Cambridge: The MIT Press.

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*, 266–297.

Rottman, B., Wk, Ahn, & Luhmann, C. (2011). When and how do people reason about unobserved causes. In P.M. Illari, F Russo, & J. Williamson (Eds.) *Causality in the sciences* (pp. 150–183). Oxford: Oxford University Press.

Rottman, B.M., & Keil, F.C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1-2), 93–125.

Saxe, R., Tenenbaum, J.B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*(1), 303.

Schlottmann, A., Allen, D., Linderoth, C., & Hesketh, S. (2002). Perceptual causality in children. *Child Development*, *73*(6), 1656–1677.

Shanks, D.R., Pearson, S.M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology Section B* (2),139–159.

Stefan, A.M., Katsimpokis, D., Gronau, Q.F., & Wagenmakers, E.J. (2022). Expert agreement in prior elicitation and its effects on Bayesian inference. *Psychonomic Bulletin & Review*, 1–19.

Stephan, S., Mayrhofer, R., & Waldmann, M.R. (2020). Time and singular causation—A computational model. *Cognitive Science*, *44*(7), e12871.

Steyvers, M., Tenenbaum, J.B., Wagenmakers, E.J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Strobl, E.V. (2019). Improved causal discovery from longitudinal data using a mixture of dags. In *The 2019 ACM SIGKDD workshop on causal discovery, PMLR* (pp. 100–133).

Tenenbaum, J.B., & Griffiths, T.L. (2001). Structure learning in human causal induction. *Advances in Neural Information Processing Systems*, 59–65.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Vidal, E., Thollard, F., de la Higuera, C., Casacuberta, F., & Carrasco, R.C. (2005). Probabilistic finite-state machines–Part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1013–1025.

Williams, J.J., & Griffiths, T.L. (2013). Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1473.

Yeung, S., & Griffiths, T.L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29.