The Role of Time in Causal Learning

Neil R. Bramley*
Experimental Psychology, UCL, London, WC1H 0DS, UK
Tobias Gerstenberg
Computational Cognitive Science Group, MIT, Cambridge, MA 02139
Ralf Mayrhofer
Institute for Psychology, University of Göttingen, Germany
David A. Lagnado
Experimental Psychology, UCL, London, WC1H 0DS, UK

Abstract

A large body of research has explored how the time course of two events affects judgments of causal strength. In this paper, we extend this work to explore the role of temporal information in causal structure induction with multiple variables. We distinguish two qualitatively different types of information: The *order* in which events occur, and the *temporal intervals* between those events. Over 3 experiments, we investigate what participants infer about the underlying causal structure of different devices after having seen their components activate over time. We focus on one-shot learning in Experiment 1. In Experiment 2, we explore how participants integrate evidence from multiple observations of the same causal device. Participants' judgments are well predicted by a Bayesian model that rules out causal structures that are inconsistent with the observed temporal order, and favors structures that imply similar intervals between causally connected components. In Experiment 3, we look more closely at participants' sensitivity to exact event timings. Participants see three events that always occur in the same order, but the variability and correlation between the timings of the events is either more consistent with a chain or a fork structure. We show for the first time that even when order cues are unavailable, people can still make accurate causal judgments on the basis of delay variability alone.

Keywords: causal; learning; structure; timing; order

Word count: 13129

*Address for correspondence:
Room 201, 26 Bedford Way
University College London
London, UK, WC1H 0DS
Tel: +44 7914419386
Email: neil.bramley@ucl.ac.uk

Introduction

Many aspects of higher level cognition, such as prediction, explanation, and goal-directed action depend on our ability to represent the causal structure of the world (Sloman, 2005; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). This makes it important to understand how people learn about causal structure as they observe and interact with the world. Research has predominantly focused on learning from trial-by-trial covariation between variables based on observations of the system (Cheng, 1997; Deverett & Kemp, 2012; Gopnik & Sobel, 2001; Perales & Shanks, 2007), and on active interventions on the system (Bramley, Lagnado, & Speekenbrink, 2015; Meder, Mayrhofer, & Waldmann, 2014; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). However, people utilize a range of sources of information in causal learning (Lagnado, Waldmann, Hagmayer, & Sloman, 2007) and human causal knowledge goes beyond mere expectations about covariation (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Sloman & Lagnado, 2015). To be able to predict and diagnose causality in real-world situations, human causal knowledge must often include beliefs about *how long* different relationships take to work (Griffiths & Tenenbaum, 2009; Lagnado & Sloman, 2006) — for example we know that turning on the heating activates the boiler almost instantly, but that the boiler will take a few minutes to make a radiator hot, and the radiator will take much longer to heat the room. Expectations about causal delays can in turn support structure inference (Buehner & McGregor, 2006; Griffiths & Tenenbaum, 2009; Hagmayer & Waldmann, 2002; Kemp, Goodman, & Tenenbaum, 2010) because the consistency between an observed event stream and the predictions of different causal models provides evidence about the underlying relationships. This temporal information is unavailable to a purely contingency-based learner. In the current paper, we focus on the role of time in causal structure induction.

Much of the extant research has focused on judgments of *causal strength*. Here, the causal structure is known and the question is how strong or reliable the relationship between cause and effect is. However, in everyday life, variables do not normally come neatly packaged as causes and effects—one must first learn the causal structure relating the variables. A number of recent papers have explored this more general problem. This work often compares human judgments with the normative predictions of Pearl's (2000) causal Bayesian network (CBN) framework. The CBN framework combines structured representations with probabilistic inference, and provides a normative calculus for reasoning both about the consequences of observations as well as hypothetical (or counterfactual) interventions. This makes CBNs suitable for generating the probabilistic action-dependent predictions we expect causal knowledge to support (Woodward, 2003).

An additional strength of the CBN framework is that, by defining a language for expressing possible models, the approach allows causal learning to be framed as a Bayesian model induction problem, where the learner uses observed evidence to infer an underlying causal structure. From this perspective, people are generally found to be effective causal learners who make inferences that are broadly normative (e.g., Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2002, 2004, 2006; Steyvers et al., 2003) but also exhibit the signatures of various inductive biases and cognitive limitations (Bramley, Dayan, & Lagnado, 2015; Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis,

2015; Mayrhofer & Waldmann, in press; Rottman & Hastie, 2013, 2016).

Strengths aside, a shortcoming of the CBN framework is that Bayesian networks do not naturally encode the temporal or spatial dimensions of causal beliefs, and so say nothing about their role in causal inference (cf. Gerstenberg et al., 2015; Gerstenberg & Tenenbaum, in press; Goodman, Tenenbaum, & Gerstenberg, 2015; Wolff & Shepard, 2013). Consequentially, many studies have focused on situations where information about time and space is non-diagnostic or abstracted away. When temporal cues have been pitted against statistical cues experimentally, judgments have tended to be dominated by temporal information (Burns & McCormack, 2009; Frosch, McCormack, Lagnado, & Burns, 2012; Lagnado & Sloman, 2004, 2006; Schlottmann, 1999). Furthermore, when researchers have tried to instruct participants to ignore event timing, participants still treated the observed timings of events to be diagnostic (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; White, 2006). These results suggest that people have strong assumptions about time's role in causality (see also Bechlivanidis & Lagnado, 2013, showing that the influence runs both ways). In the current paper we take a novel approach: we eliminate statistical contingency information, and focus exclusively on what participants can learn about causal structure from temporal information alone.

Structure of the paper

The structure of the paper is as follows. First, we review the literature on causal learning and time. After describing the learning problem we focus on in the current paper, we then outline the Bayesian framework and methodology used to explore human causal learning in time. Our first experiment explores one-shot causal structure judgments, based on a single observation of a simple device operating through time. In Experiments 2 and 3, we look at how people integrate evidence from multiple clips of the same device. In Experiment 3, we focus on temporal delays in a situation where there is no temporal order cue: three events always occur in the same order, but the variability and correlation between the timings of the events is either more consistent with a chain or a fork structure. Finally, we discuss the scope of our findings and propose future research.

Existing research

Temporal information is relevant for causal judgments in at least two ways. The *temporal order* of events is important since causes cannot precede their effects. The *timing* of events provide additional information, since they are diagnostic about the true underlying causal structure when multiple possible structures are consistent with the observed temporal order.

Temporal order. The assumption that causes precede effects is at the heart of our notion of causality (Hume, 1740) meaning that, prima facie, the order in which events occur is a highly important cue to causality. Inference from perceived order appears to be natural, almost automatic. For example, Wolff and Shepard (2013) cite multiple reports, following a 1997 power blackout in New York, of people having the sensation that an action they had taken just before the blackout (touching a doorknob, plugging in an appliance, jamming a ceiling fan) was its cause. Magicians use this to trick their audiences into believing they can

affect objects at a distance, snapping their fingers just before revealing their masterstroke (Kuhn, Caffaratti, Teszka, & Rensink, 2014).

Precedence also forms the basis for many legal judgments, with establishment of the order of the events in a case often playing a large role in attribution of responsibility for a crime (Lagnado, 2011; Lagnado & Gerstenberg, in press). Additionally, an important concept in economics, *Granger causality* (Granger, 1969), uses the extent to which past values of one variable can be used to predict current variation in another as a marker for causation.

Rottman and Keil (2012) explored causal induction in situations where variables were measured at discrete intervals. For example, one might measure barometric pressure and precipitation on successive days. Finding that barometric pressure was high on Monday and Friday and it rained on Tuesday and Saturday invites the inference that high pressure causes rain. In seven experiments, the authors find that people readily attribute causal relationships from variables that changed state at time $t-1$ to those that changed state at time t, and do so even when a cover story suggests there should be sequential independence. They argue that people's default representation of causality is as a qualitatively ordered sequence of changes, and suggest that estimating statistical dependence across multiple independent instances, as in contingency-driven structure learning, is a more difficult, less natural mode of causal reasoning.

Experienced event order also affects people's causal judgments when events take place in continuous rather than discretized time. Lagnado and Sloman (2006) explored a situation that contrasted trial-by-trial covariation with temporal order cues. In their experiment, a virus propagates through a computer network causing computers to be infected with different temporal delays. Participants' task was to infer the structure of these computer networks based on having observed the virus spreading through the network multiple times. Participants preferred causal models that matched the experienced order in which computers got infected, even when trial-by-trial covariation cues went against temporal order cues.

Several studies further suggest that people are reluctant to endorse causal connections between events which appear to occur at the same time. Burns and McCormack (2009) found that by age 6 to 7, children strongly favor a $B \leftarrow A \rightarrow C$ common cause over a $A \rightarrow B \rightarrow C$ chain when they observe B and C happening simultaneously and after A. Even when the causal mechanism is plausibly instantaneous (as in Frosch et al., 2012; Lagnado & Sloman, 2006; McCormack et al., 2016), people tend to attribute simultaneous activations of components to a common cause. However, previous work has not looked at what people infer from situations in which observed simultaneity cannot be attributed to a common cause, but must have either be instantaneous or coincidental. Additionally, little work has looked specifically at how people integrate evidence of events occurring in different orders over multiple trials (although see Lagnado & Sloman, 2006).[1]

Event timings. Going beyond temporal order, we can also consider the exact timing of events as another source of information about causal relationships (Hagmayer & Waldmann, 2002). Using *only* temporal precedence to guide judgments would put everything that ever happened on equal footing as a candidate cause; a switch you switched a

[1]We use the $\rightarrow$ operator to denote a causal relationship between events (e.g., $A \rightarrow B$ means A caused B), and $\succ$ operator to denote event *order* (e.g., $A \succ B$ means A preceded B).

year ago would be just as likely a cause of a light turning on, as one you just switched.[2] In this section we discuss what role event timing plays in guiding causal judgments.

In the associative tradition, causal relationships are treated as another form of learned association, where the constant conjunction, and temporal as well as spatial contiguity between two variables naturally leads to their being associated by the cognitive system (Hume, 1748/1975). Since increased intervals rapidly reduce the rate of associative learning (Grice, 1948; Shanks & Dickinson, 1987; Wolfe, 1921), associative theories generally predict that judgments of causality will show this same pattern. Making similar predictions, early cognitive theories (Ahn, Kalish, Medin, & Gelman, 1995; Einhorn & Hogarth, 1986) suggested that the more distant two events are in time, the more costly it will be to sustain the first event in working memory long enough to relate it to the second event, leading to monotonic reduction in causal judgments. Lagnado and Speekenbrink (2010) identify an additional normative reason for why delays will often lead to reduced judgments of causality. All things being equal, the longer the gap between putative cause and effect, the more likely it is that other events may have occurred in the meantime that could have also caused the effect.

Humans are able to make causal inferences that are sensitive to expectations about event timing. When participants are given information about causal mechanisms that imply different delays, their resultant causal judgments are strongly influenced by expectations about average delay length and variability of the mechanisms. Seeing shorter-than- as well as longer-than-expected intervals leads to reduced judgments of causality (Buehner, Cheng, & Clifford, 2003; Buehner & May, 2002, 2004; Greville & Buehner, 2010; Hagmayer & Waldmann, 2002; Schlottmann, 1999). For example, seeing a regular light bulb come on several seconds after switching a switch is rated as less causal. However, the case is different if you learn that it is an energy saving bulb which takes time to warm up.

As well as unexpected time intervals, variability in intervals across trials has been shown to reduce judgments (Greville & Buehner, 2010; Greville, Cassar, Johansen, & Buehner, 2013; Lagnado & Speekenbrink, 2010). However, these studies have focused on situations in which there is a single candidate cause–effect pair. In this paper, we explore the more general problem of inferring the causal structure of multiple variables based on observations of events in time.

Neuroimaging data also supports the idea that timing expectations play a role in causal learning (Jocham et al., 2016) . In two behavioral experiments, participants' task was to identify occasional rewarding events in event streams. The results showed that both associative and "contingent" — or theory-dependent — learning take place simultaneously and in separable brain circuits — the former predominantly in the amygdala, and the latter in the orbitofrontal cortex. Amygdala learning was associative in the sense that it learned relations between rewards and preceding events irrespective of task instructions. Orbitofrontal learning was contingent in the sense that it depended dynamically on instructions about what delays to expect for genuine stimuli and reward relationships, only attributing rewards to appropriately timed stimuli. A central goal in causal learning research is to understand where these theory-dependent judgments come from. How are people and animals often able to make strong and sensible "one-shot" inferences about

[2]Of course, there are more intervening events in the former case, providing another possible avenue capturing why the latter is a better candidate (e.g. Lagnado & Speekenbrink, 2010).

causal structure, without explicit instruction, in situations where naive statistical learning algorithms would require much more data?

Several researchers have suggested that causal theories might underpin such "one-shot" inferences (Goodman, Ullman, & Tenenbaum, 2011; Griffiths, 2005; Griffiths & Tenenbaum, 2009; Kemp et al., 2010). The idea is broadly that, over the course of development, people organize their causal general knowledge hierarchically, with the core abstract features of causation at the top and increasingly domain- and context-specific features below. Each level of the theory generates a probability distribution on variables at the level below, and the more specific the subdomain, the greater the constraints on the space of possible hypotheses. As a learner's world knowledge gets richer, their causal judgments can rely more strongly on identification of the right domain and application of domain-specific knowledge and constraints, resulting in apparent "one-shot" inferences (see also Lake, Salakhutdinov, & Tenenbaum, 2015).

The theory-based causal inference framework provides an explanation for the role of temporal expectations in causal induction. By learning the typical cause–effect delays in a particular domain, a learner can use this knowledge to rapidly identify new connections when candidate events are appropriately spaced in time. Griffiths (2005) showed how different expectations about delay distributions allow for strong one-shot inferences about a causal process. In their experiments, participants made causal judgments about "nitroX" barrels that were causally connected and exploded in different sequences. Because different causal models imply very different event timings, the Bayesian model was able to rapidly infer the causal structure from an observed sequence of exploding barrels. Building on this work, Pacer and Griffiths (2012) model causal influence in situations where a discrete event affects the rate of occurrence of another variable in continuous time. In particular, they capture people's judgments about the causal strength of variables that affect the rate of bacteria death in a population over a number of days (cf. Greville & Buehner, 2007). Extending this approach, Pacer and Griffiths (2015) also capture inferences about relationships for which the influence of the cause on the effect is expected to last for some time before it gradually dissipates. Using this model, Pacer and Griffiths explain participants' inferences about which of three occasionally occurring seismic waves affected the rate of occurrence of earthquakes (see Lagnado & Speekenbrink, 2010, Experiment 2). As predicted by the model, participants' judgments were affected by uncertainty about the number of intervening events rather than the absolute intervals between putative causes and effects.

Pacer and Griffiths' approach is well-suited to capturing situations where events alter the *rate* of occurrence of other events. However, it does not readily apply to situations in which causes bring about their effects exactly once. For example, an event at A in their model might increase the *number of activations* of B that you expect to see over the next 5 seconds from 0.1 to 1.1. However in their representation the number of events that occur in total is inherently stochastic. This means that the occurrence of the cause might sometimes result in no activation of B, or in several activations of B. In this paper, we are interested in situations in which the causal relation between two events is singular — that is, the cause affects the effect exactly once.

In summary, research has established that temporal information plays an important role in how people make causal judgments (Lagnado & Sloman, 2006; Rottman & Keil, 2012; Sloman, 2005). Causal inference seems to be driven by temporal information partly

via automatic (Michotte, 1946/1963) and developmentally basic (Burns & McCormack, 2009) mechanisms, but also through more complex theory-contingent modes of thinking (Griffiths & Tenenbaum, 2005). While previous work has explored representation of causality in time (Griffiths, 2005; Pacer & Griffiths, 2012), no research to date has proposed a model that is sensitive to temporal order and incorporates expectations about intervals between particular events.

Modeling causal induction from temporal information

Despite the wealth of research on time and causal learning, temporal information has not been subsumed into a unifying framework for understanding how causal beliefs are formed to the same extent as contingency information has. In this section, we lay out a learning problem that isolates the role of temporal information. We then present our Bayesian approach to modeling learning in this situation, distinguishing learning based on information about *temporal order* alone from learning based on forming parametric expectations about *temporal intervals* between causes and effects.

The learning problem

To isolate temporal information, we focus on situations where the learner must identify the causal structure of a system made up of a number of components that are causally related but in which the causal links take time to propagate. We assume that the causal relationships are known to be generative and sufficient in the sense that the activation of a cause component will invariably lead to the activation of its effect component(s), but where the delays between activation of the cause and the effect are variable across instances. In Experiments 1 and 2 we focus on judgments about the causal structure of a simple system with two causal components A and B and an effect component E that form a hypothesis space of seven possibilities (Figure 1). In Experiment 3, we will focus on a more restricted space with a single cause component S and two components A and B that are either its direct or indirect effects. Evidence in all the experiments consists of clips that show how the different components of a causal device activate over time. In Experiments 1 to 2 participants are told that the parentless components in the device activate due to background causes while in Experiments 3 the learner activates the system themselves.

The different causal connections of a causal structure might exhibit different delays — for example, in the A-fork (see Figure 1) it might take longer on average for A to cause B than for A to cause E. Furthermore, the same connection might also exhibit variability in delays across trials — for example, A's causing E might be subject to longer or shorter delays on different occasions. As a consequence of this variability, many causal structures can generate several qualitatively different *orders* of activation.

Bayesian models of learning

From a Bayesian perspective, learning is the process of updating a probability distribution over the true state of the world, where the ground truth is treated as a random variable and its possible values make up the hypothesis space. A Bayesian learner updates their prior probability distribution into a posterior distribution as evidence is observed. The posterior from one learning instance becomes the prior for the next, and this process

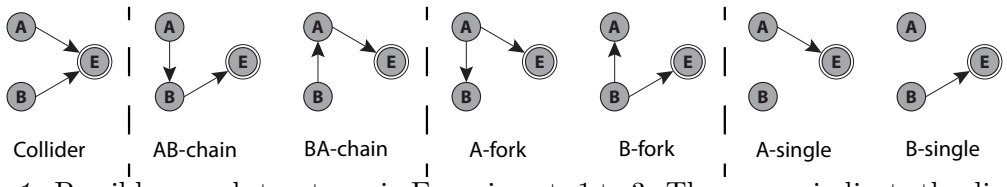

Figure 1. Possible causal structures in Experiments 1 to 3. The arrows indicate the direction of the causal relationship. Dotted lines indicate different types of structure. *Note*: the Collider is conjunctive — both A and B must occur for E to occur.

continues as evidence is received. With sufficient evidence, the learner's subjective beliefs eventually approximate the ground truth provided that the hypotheses are distinguishable and the hypothesis space contains the ground truth

Exact Bayesian inference is intractable for most realistically complex problems. However, for a suitably constrained problem space like the one explored here, Bayesian inference provides a powerful framework for understanding human learning. We can look at how people update their beliefs as evidence is presented, and learn about the prior assumptions they bring to the task.

In the current context, the random variable we are interested in is the true underlying causal structure $s \in \mathcal{S}$ in the set of possible structure hypotheses $\mathcal{S}$, and data will take the form of n observed patterns of component activations over time $\mathbf{d} = (d_1, d_2, \ldots, d_n)$. We update a prior belief about the possible underlying structures $p(\mathcal{S})$ to a posterior belief over the structures given the data $p(\mathcal{S}|\,\mathbf{d})$ using Bayes theorem

$$p(\mathcal{S}|\,\mathbf{d}) \propto p(\mathbf{d}\,|\mathcal{S}) \cdot p(\mathcal{S}), \tag{1}$$

where $p(\mathbf{d}\,|\mathcal{S})$ is the likelihood function over structures $\mathcal{S}$.

For inference to proceed, the learner needs a likelihood function determining how likely each structure would be to exhibit the set of experienced temporal patterns $\mathbf{d}$. We first propose a class of models based on simple likelihood functions that ignore the exact timing of events but assign likelihoods simply based on their temporal ordering. We consider two models that differ in whether they allow for instantaneous causation, that is, causes and effects happening at the same time. We then consider a richer framework that incorporates expectations about causal delays. We show how, based on the principles of Bayesian Ockham's razor (MacKay, 2003), both approaches form preferences for different causal structures requiring neither contingency information nor pre-existing expectations about the duration or variability of the delays.

Only order matters

Likelihood functions. The order of events constrains what structures are capable of having produced the observed evidence. We capture the information contained in the temporal order of events in a simple model that divides its likelihood evenly across all order-consistent patterns. Hence, any particular sequence of component activations has likelihood $1/N$, where N is the number of distinct temporal orderings consistent with that structure (Figure 2b and c, columns). In the following, we use the $\succ$ operator to denote event order. For example, $A \succ B \succ E$ means that A preceded B which preceded E. $AB \succ E$ means that A and B happened simultaneously before they were succeeded by E.

In the A-fork, A is the cause of both B and E, therefore this structure is consistent with patterns in which A preceded both B and E ($A \succ B \succ E$, $A \succ E \succ B$ and $A \succ BE$, see Figure 2a) but inconsistent with any pattern where either B or E precede A. Whether $AB \succ E$ or $AE \succ B$ are consistent with the A-fork depends on whether one assumes causes and effects can occur simultaneously. In order to test whether people make this assumption, we will compare two variants of our model to participants' judgments. *Order non-simultaneous* (Order_{NS}) makes the non-simultaneity assumption meaning only events that strictly precede other events are candidate causes. *Order simultaneous* (Order_{S}) relaxes this assumption, such that an event can be the cause of another event even if they occur at the same time. For Order_{NS}, the AB-chain is only consistent with $A \succ B \succ E$ (Figure 2b, second column). For Order_{S}, the AB-chain is also consistent with $AB \succ E$ and $A \succ BE$, and thus this model variant spreads its likelihood more widely.[3]

Because some structures are compatible with fewer kinds of evidence patterns than others, the order models will tend to favor them over a more flexible structure that can also produce the evidence seen. For example, under the non-simultaneity assumption, pattern $A \succ B \succ E$ in row 2) is the only pattern consistent with the AB-chain, while A-single is consistent with all but two types of patterns and thus spreads its likelihood much more widely. Switching focus from Figure 2's columns to its rows gives a perspective on the models' posterior predictions. For instance, upon observing a device that activates in the $A \succ B \succ E$ order, Order_{NS} will favor the AB-chain, even though it has not ruled out the Collider, the A-fork, or either of the two single-link structures A-single and B-single.

As another example, after observing pattern 1) $AB \succ E$, the Order_{NS} model will rule out all structures except for the Collider, A-single, and B-single. Between these remaining structures, it prefers the Collider since it is consistent with fewer types of pattern. In contrast, the Order_{S} model cannot rule out any structure based on this evidence. It has a slight preference for the AB-chain and the BA-chain since these two structures are compatible with the fewest number of different temporal order patterns.

Inference. After seeing data $\mathbf{d}$ in the form of one or several temporal order patterns, inference proceeds by updating a prior over causal structures $\mathcal{S}$ to incorporate this data.

The order models only consider the qualitative ordering of the component activations, for example $\mathbf{d} = (d_1 = \{A \succ B \succ E\}, d_2 = \{AB \succ E\}, \ldots)$, where d_i indexes independent observations of the device. The models yield various posterior beliefs based on different sequences of temporal activation $\mathbf{d}$. For example, starting from a uniform prior over the seven structures, Figure 3 shows posteriors under the simultaneous and non-simultaneous assumptions based on having observed three patterns of activation d_1, d_2, and d_3, and then again after having observed a fourth pattern d_4. In the first example (top row), both non-simultaneous and simultaneous models favor the Collider after d_1, d_2, and d_3 and their preference increases with d_4. In the second example (middle row), both models prefer the

[3]Note that several additional possible patterns are not pictured: $AE \succ B$, $BE \succ A$, and ABE. We do not use these in our experiments because, by being inconsistent with all structures under the non-simultaneity assumption their appearance would force a simultaneity assumption on participants. However, if people actually make the simultaneity assumption, $AE \succ B$ and $BE \succ A$ are each consistent with one fork and one single, and ABE is consistent with all seven structures. Thus we divide the Order_{S} likelihoods across these additional structures, yielding columnwise likelihoods of $[\frac{1}{6}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{8}, \frac{1}{8}]$. We excluded any patterns in which E alone occurred first, since we instructed participants that E is always caused by either A, B, or both.

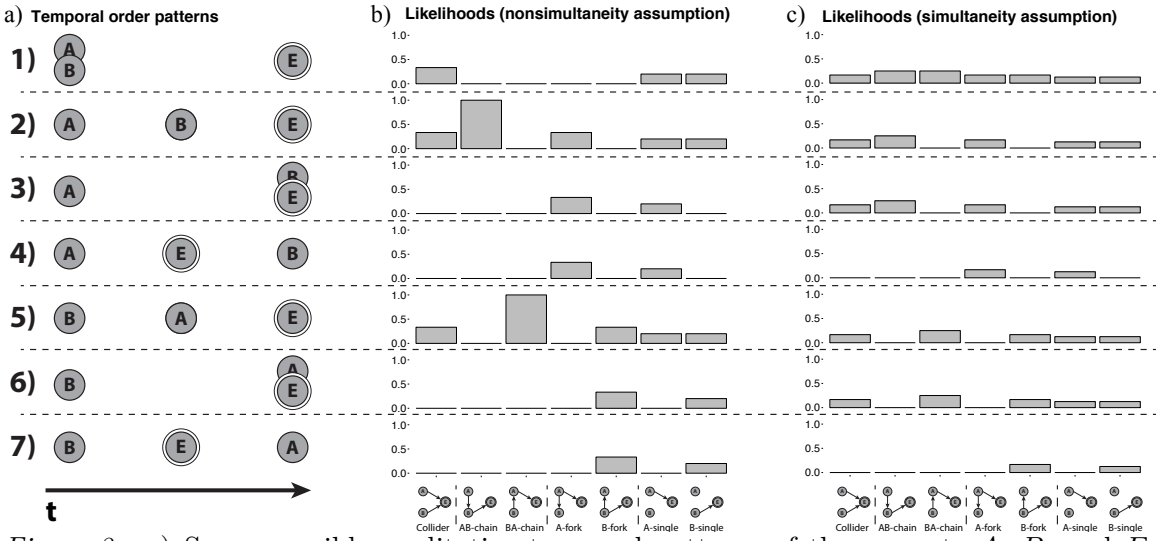

Figure 2. a) Seven possible qualitative temporal patterns of three events A, B, and E. Likelihood functions for the pattern types given the seven different causal structures with non-simultaneity assumption b) or simultaneity assumption c).

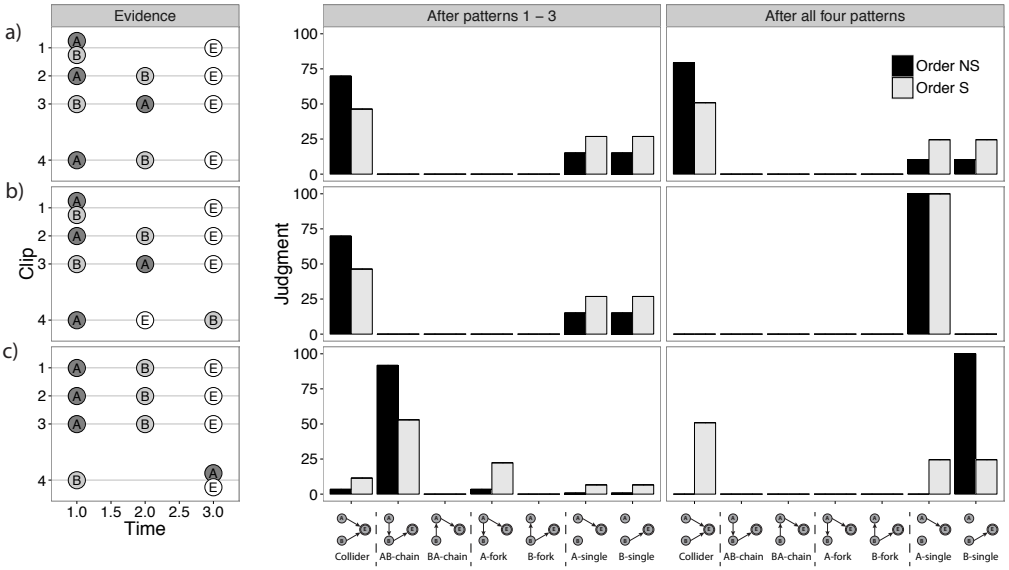

Figure 3. Three examples of order model predictions. Left hand side: Sets of 4 time series showing staggered activation of components A, B and E. Right hand side, model posteriors after seeing clips 1–3 (left column), and after having seen all four patterns (right column).

Collider after d_1, d_2, and d_3 but switch upon d_4 which rules out out all the structures except the A-single. In the third example (bottom row), the two assumptions lead to quite different predictions, with the non-simultaneous model preferring the Collider and the simultaneous model preferring the B-single after seeing all the clips.

Timing matters

Generative model. The order models make the strong assumption that any activation pattern whose temporal order is consistent with the device is equally likely. While simple to work with, the assumption is inconsistent with more specific beliefs about delays between causes and effects. For example, people may believe that causes take a certain amount of time to bring about their effects and that these delays will be similar across instances. To capture these intuitions, we need richer models than Order_{NS} and Order_{S} – models that incorporate assumptions about how the events and their timings are being brought about.

In our task, an observed temporal pattern d_i consists of the activation times t_X of the three components A, B, and E; thus $d_i = \{t^i_A, t^i_B, t^i_E\}$. We will use t_{XY} to refer to the temporal interval between the activations of X and Y (i.e., $t_{XY} = t_Y - t_X$). Additionally, we will use $t_{X\to Y}$ to distinguish causal delays from temporal intervals t_{XY} which are not necessarily causal.

Independent causes. We start with formalizing the timing of independent causes which do not have any parents in the causal structure (such as variables A and B in the Collider, A-single, and B-single). Analogously to causal Bayes nets, in which independent causes are assumed to be statistically independent of each other (i.e., uncorrelated), we define independent causes to be temporally independent of each other as well as independent from the (artificially determined) beginning of the clip. The natural candidate for modeling such events is the exponential distribution, which is is "memory-less". This property means that how long you expect to wait for an event is independent of how long you have already been waiting for it. Thus, the information that another (independent) event has happened does not alter your expectation about the activation time of the next event. If X is an independent (i.e., parentless) cause then the timing of X is determined by

$$p(t_X|\lambda) = \lambda e^{-\lambda t_X} \tag{2}$$

with $p(t_X|\lambda) = 0$ for activation times smaller than 0 and expectation $\frac{1}{\lambda}$.

Causal links. The generalization of the exponential distribution is the gamma distribution. It introduces time dependence, and it is therefore the natural candidate to model the relative timing of causally related events. Gamma distributions can be defined by a shape parameter α and an expectation μ. Under the assumption that X causes Y, the timing of Y depends upon the timing of X such that $t_Y = t_X + t_{X\to Y}$ with $t_{X\to Y}$ being gamma distributed:

$$p(t_{X\to Y}|\alpha,\mu) = \frac{(\frac{\alpha}{\mu})^\alpha}{\Gamma(\alpha)}(t_{X\to Y})^{\alpha-1}e^{-\frac{\alpha}{\mu}t_{X\to Y}} \tag{3}$$

with $p(t_{X\to Y}|\alpha,\mu) = 0$ for temporal delays smaller than 0 (i.e., no backwards causation).

Figure 4 shows examples of gamma distributions for different parameter values. The gamma distribution is flexible and allows to represent a continua of short (small μ) to long (large μ) and variable (low α) to reliable (high α) delays.

As $\alpha \to \infty$, the gamma distribution becomes increasingly centered around its expected value, capturing what we will call "positive" time dependence (e.g., Figure 4, solid and dashed lines). One's expectation about the time of an effect increases following the observation of its cause, peaking around its mean and then dropping away again. For $\alpha = 1$

the gamma distribution is an exponential distribution. Values of $\alpha < 1$ capture "negative" dependence whereby upon observing a cause one expects to see its effect either right away or in a very long time (e.g., Figure 4, dot-dashed line).

Colliders/Common-effect structures. Within this framework, the Collider (i.e., common-effect structure) presents a special modeling challenge since it involves a joint influence of two distinct causes. There are various plausible combination functions for capturing this kind of joint influence. We explicitly stipulate in all experiments that the Collider structure is conjunctive, meaning that the activation of E occurs only after the activations of both A and B and, by implication, the arrival of both of their causal influences at E. To model this, we consider the t_E in a Collider structure to be the maximum of the two unknown causal delays for $t_{A\to E}$ and $t_{B\to E}$ offset by their activation time

$$t_E = \max[t_A + t_{A\to E}, t_B + t_{B\to E}] \tag{4}$$

with $t_{A\to E}$ and $t_{B\to E}$ being gamma distributed (see Equation 3) and t_A and t_B being exponentially distributed events (see Equation 2).[4] Note that a *disjunctive* Collider is modeled by simply using the minimum instead of the maximum in Equation 4 (see Equations A-4 and A-6 in the Appendix).

Likelihood functions. The generative model laid out above provides the formal tools we need to determine the likelihood of any observed temporal pattern given a structure hypothesis. To distinguish different causal structures, we translate the absolute timings of a set of events into specific cause–effect pairings, depending on the parents $\mathrm{pa}(X)$ of each variable under the structure at hand. For instance, absolute timings $\{t_A, t_B, t_E\}$ will be translated into $\{t_A, t_{AB}, t_{BE}\}$ with $t_{AB} = t_B - t_A$ and $t_{BE} = t_E - t_B$ under the AB-chain hypothesis. Dependent on different beliefs about the underlying causal structure and delay distributions, the same set of observed activation times will be more or less likely as we will illustrate below.

Sometimes it may be reasonable to assume that the different connections in a causal system have the same underlying delay distribution (e.g., they might all be components of the same type). In other situations, we might expect completely different delays for different parts of a process (for example it might take millions of years for the wind to wear through a rock face but only seconds for the freed rock to fall and cause a landslide). We can embody these different assumptions with different model variants. The *pooled model* (Delay_P, Figure 5a) has a single α and μ parameter for all the delays within a single structure $s \in \mathcal{S}$. In contrast, the *independent model* (Delay_I, Figure 5c) has separate parameters α_c and μ_c for each causal connection $c \in C_s$ where C_s is the list of all connections in structure s. To capture weaker assumptions (e.g., that the delay distributions for relationships within a device are related but not identical), one could extend this with a *hierarchical model* (Delay_H, Figure 5b) that combines expectations about the variability of the different distributions within a device via hyperparameters that define distributions for α and μ, although we do not do this in the current paper.

We start by describing the likelihood function of the *pooled* Delay_P variant of the model. The likelihood of a temporal pattern d_i given a causal structure $s \in S$ with timings

[4] We derive the full equations for the Collider likelihood assuming shared parameters for the input connections (as in Delay_P) and separate parameters (as in Delay_I), in the Appendix.

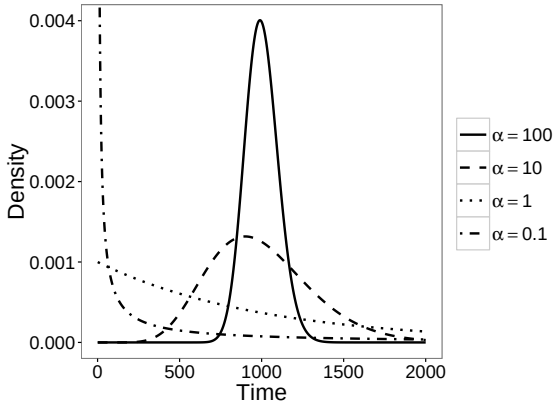

Figure 4. Three example gamma distributions. All have a mean of $\mu = 1000$ ms but differ in their shape α. The exponential distribution is the case where $\alpha = 1$.

governed by parameters λ, α and μ, is the product of the likelihoods of the relative delays between causes and effects that result from mapping the absolute event timings $t_X \in d_i$ onto the structure of the model $s \in S$

$$p(\mathbf{d}\,|\lambda, \alpha, \mu; s) = \prod_{i \in 1:n} p(d_i|\lambda, \alpha, \mu; s) = \prod_{i \in 1:n} \prod_{t_X^i \in d_i} p(t_X^i - t_{\mathrm{pa}(X)}^i|\lambda, \alpha, \mu; s) \tag{5}$$

with $p(t_X^i - t_{\mathrm{pa}(X)}^i|\lambda, \alpha, \mu; s)$ being either gamma or exponentially distributed (see Equation 3 and Equation 2, respectively) depending on whether X has a parent or not (and assuming that the structure is not a Collider).

For the Collider, we have to determine the joint likelihood of t_{AE} and t_{BE}. Note that we use "$\rightarrow$"s to distinguish the unknown true delays from the observed inter-event intervals since either t_{AE} or t_{BE} may include some time spent "waiting" for the other causal influence to arrive. Observed event timings depict one of two mutually exclusive cases: either the causal influence of A arrived later (i.e., $t_{A\rightarrow E} = t_{AE}$) overshadowing the timing of B's causal influence (i.e., $t_{B\rightarrow E} \leq t_{BE}$) or the causal influence of B arrived later (i.e., $t_{B\rightarrow E} = t_{BE}$) overshadowing the influence of A (i.e., $t_{A\rightarrow E} \leq t_{AE}$). The joint likelihood of the observed intervals are then given by the sum of their individual likelihoods

$$p(t_{AE}, t_{BE}|\alpha, \mu) = p(t_{AE}|\alpha, \mu) \cdot p(t_{B\rightarrow E} \leq t_{BE}|\alpha, \mu) + p(t_{BE}|\alpha, \mu) \cdot p(t_{A\rightarrow E} \leq t_{AE}|\alpha, \mu) \tag{6}$$

with $p(t_{AE}|\alpha, \mu)$ and $p(t_{BE}|\alpha, \mu)$ being gamma distributed (see Equation 3), and $p(t_{A\rightarrow E} \leq t_{AE}|\alpha, \mu)$ and $p(t_{B\rightarrow E} \leq t_{BE}|\alpha, \mu)$ following the cumulative distribution function of the gamma distribution (i.e., the integral over Equation 3 with upper bound t_{AE} or t_{BE}, respectively; see Appendix for a more detailed derivation).

In the general case, λ, α, and μ are unknown. To get the (marginal) likelihood of the data given the structure, which is our target for Equation 1, we have to marginalize out the parameters by integration, assuming some prior distribution over λ, α, and μ[5]

[5]Concretely, we used an Exponential (0.1) prior for α, an Exponential (0.0001) prior on μ and an Exponential (10000) prior on λ, corresponding to a weak expectation for positive dependence, shorter delays and frequently occurring independent causes.

$$p(\mathbf{d}\,|s) = \int p(\mathbf{d}, \lambda, \alpha, \mu|s)\ d\lambda\ d\alpha\ d\mu \tag{7}$$

$$= \int p(\mathbf{d}\,|\lambda, \alpha, \mu; s) \cdot p(\lambda, \alpha, \mu|s)\ d\lambda\ d\alpha\ d\mu \tag{8}$$

$$= \int p(\mathbf{d}\,|\lambda, \alpha, \mu; s) \cdot p(\lambda|s) \cdot p(\alpha|s) \cdot p(\mu|s)\ d\lambda\ d\alpha\ d\mu \tag{9}$$

We discuss how we approximated these integrals and sensitivity to priors in the Appendix.

To see how this timing sensitivity supports causal structure inferences, let us assume that a learner observed the following order of activation: $A \succ B \succ E$. If they make the Delay$_P$ assumption that cause–effect delays for the connections in this device come from the same distribution, we would expect their belief about whether the underlying causal structure was a Collider, an AB-chain, or an A-fork to shift depending on t_{AB} and t_{BE}. Intuitively, if t_{AB} and t_{BE} are similar, this seems most consistent with an AB-chain. However, if t_{BE} is very small this seems more consistent with the A-fork (in which $t_{A\to B}$ and $t_{A\to E}$ would be similar). If t_{AB} is very small then the device might be a Collider (where we would expect $t_{A\to E}$ and $t_{B\to E}$ to be similar). Delay$_P$ makes these predictions via Bayesian Occam's razor. Essentially, it assumes all causal delays of the connections in a device follow the same underlying gamma distribution $G_{X\to Y}(\alpha, \mu)$. Even if we have only a vague idea what specific form this distribution takes (as specified by α and μ), the model will still tend to favor whatever causal hypothesis renders these causal event timings the most similar on average. The more tightly clustered the inferred delays are, the more compact the generative causal delay distribution can be (here a high average α parameter), which leads to higher likelihoods assigned to the data points. See Figure 5a for an illustration of this point.

Inference in the *independent* Delay$_I$ model (Figure 5c) proceeds in same way, but with separate parameters for the delay distributions of the different causal connections $c \in C$ [e.g., $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{|C|})$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{|C|})$]. That is, it assumes there is no relationship between the delays of different parts of a causal device. The distribution of delays implied by mapping event timings onto different causal models can still be diagnostic, provided one interacts with the same device more than once. Figure 5c gives an illustration of this. Here, the temporal intervals t_{SB} are consistently around 2 s, while t_{SA} and t_{AB} are much more variable. We can explain these patterns of evidence more parsimoniously by assuming that the true structure is an S-fork with a regular $S \to B$ connection and an irregular $S \to A$ connection. It is not impossible that the true structure is a chain, but the chain structure cannot explain the additional systematicity in the data whereby the $t_{S\to A}$ and $t_{A\to B}$ intervals almost perfectly cancel out.[6]

Summary

In summary, the non-simultaneous and simultaneous order models (Order$_{NS}$ and Order$_S$) operationalize inference based purely on the qualitative ordering of observed activations. These models show how certain structures can be ruled out, and some of the

[6]We note though that with additional assumptions about the functioning of the device the reverse inference might hold. For example, if the $A \to B$ connection was somehow designed to cancel out variation in $S \to A$ so as to lead to a reliable t_B.

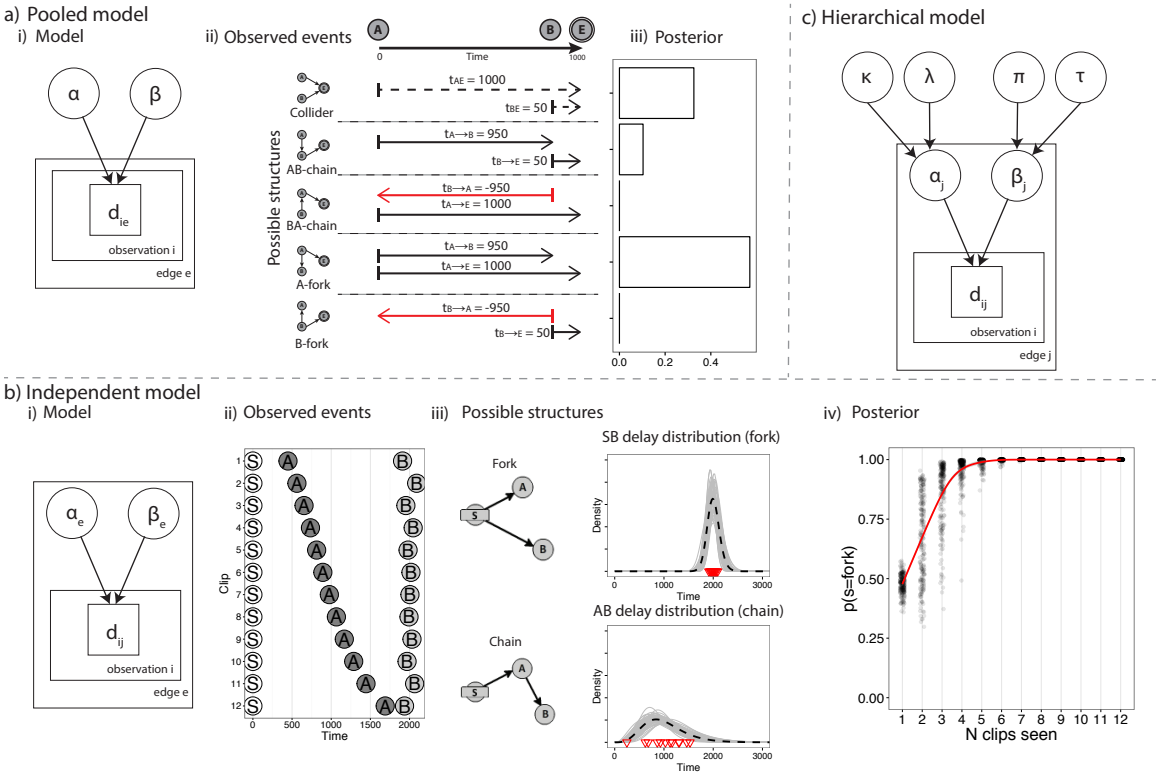

Figure 5. Delay sensitive models and predictions. a) i. *Pooled* Delay_P in plate notation. ii. Example of inference in the pooled model. Observed event timings are mapped onto causal delays under different models. Each row shows the causal delays assuming a different structure. For the Collider, dashed lines indicates that one or other causal delay may be shorter than the observed intervals. Red arrows indicate that structures that can be ruled out based on order alone. iii. Posterior predictions of the delay model assuming priors of $S \sim \text{Unif}(\frac{1}{7})$, $\alpha \sim \text{Exp}(0.1)$, and $\mu \sim \text{Exp}(0.0001)$. b) i. *Independent* Delay_I model in plate notation. ii. 12 patterns of evidence. iii. Posterior marginal inference for two possible structures. The plots show posterior delay samples (gray lines) and their overall density (dotted black line). Both structures share the same $t_{S\to A}$ delays, but the high variance of t_{AB} relative to t_{SB} means this data was more likely produced by a fork as shown in iv., which plots the posterior probability of the fork structure averaged over subsets of the 12 clips (red line gives smoothed average, black dots give posteriors for samples). c) An example of a *hierarchical* Delay_H model in plate notation, where different components have different distributions but are related by hyperparameters.

remaining structures preferred, based on the order of events alone. What structures are ruled out depends on whether simultaneous events are considered consistent with causation. While these order-based models are good at ruling out inconsistent causal structures, they are limited in their ability to distinguish between structures that are consistent with the observed order of events.

The delay-based models *pooled* Delay_P, *independent* Delay_I and *hierarchical* Delay_H make inferences within the space of hypotheses not-yet-ruled-out by Order_{NS}, but distribute

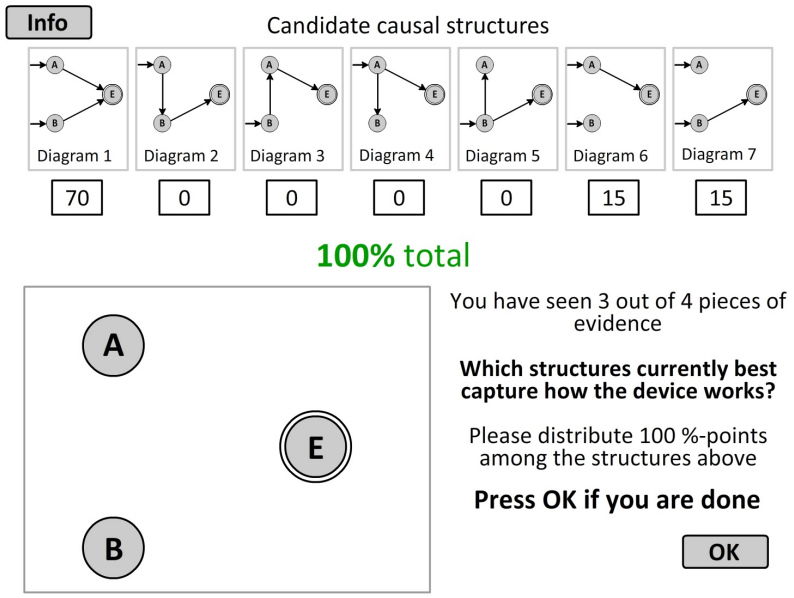

Figure 6. The experiment interface for Experiments 1–3. Clips are shown in the bottom left panel and judgments elicited at the top.

their likelihood very differently depending on the expected rate and variability of the various inter-event intervals. Assuming an uninformative prior on shape α and mean μ, the pooled delay model Delay_P favors whichever structures render the experienced inter-event intervals the most regular across all connections and all instances, while the independent delay model Delay_I favors whichever structures imply the most regular within-edge delays on average, even if these differ considerably for different connections. A hierarchical model would make predictions somewhere in between, allowing that different connections can have different delays but that they are still related.

Overview of Experiments

The task

We designed a task environment in which participants observed causal devices exhibiting one or several patterns of activation, and then made judgments about how they thought the components of that device were causally connected. Evidence was presented in the form of short movie clips. Each clip simply showed three components, (A, B, and E in Experiments 1–2, and S, A and B in Experiment 3), which were represented by circles and arranged in a triangle (see Figure 6 bottom left). During each clip, all three components activated by turning from white to gray (Experiments 1–2) or from white to yellow (Experiment 3). Activated components remained colored until the end of the clip. To minimize people's context-specific expectations about what causal structures or delays were more likely a priori, we kept the task abstract. Participants were not told anything about what kinds of causal processes underlie the activation of the different components.

Possible causal structures. As discussed in the introduction, we restricted the space of possible causal structures to seven in Experiments 1–2 (see Figure 1) and two in

Experiment 3. In Experiments 1 and 2, each structure featured two candidate causes A and B and one effect E. Participants were informed that the Collider structure is conjunctive, meaning that both A and B must activate in order for E to occur. In Experiment 3 there was a starting component S and two candidate effects A and B. The true structure was either a chain (e.g., $S \rightarrow A \rightarrow B$) or a fork (e.g., $A \leftarrow S \rightarrow B$).

Eliciting judgments. In order to have a fine-grained measure of participants' beliefs, we asked participants to distribute 100 percentage points over the set of possible candidate causal structures, such that each value indicated their belief that the given structure is the one that generated the observed evidence (see Figure 6 top). We can then directly compare participants' distributions over the structures with the predicted posterior distributions based on our different models.

Experiment 1: One-shot inferences

In Experiment 1 we explored one-shot inference. We asked participants to make judgments about causal devices after watching a single clip and replaying it several times. We varied the timing and order of the activation of the three components systematically across problems. Depending on whether or not participants rule out instantaneous causation, we expected the judgments to better match the predictions of the non-simultaneous or simultaneous order model, respectively. If participants' judgments were, in addition to temporal order, also sensitive to timings, we expected them to assign more points to structures that imply similar cause–effect delays (e.g., a fork if B occurs very early as in clip 2 shown in Figure 7a, and a Collider if B occurs very late as in clip 6).

Methods

Participants and materials. Thirty-one participants (18 female, $M_{age} = 36.8$, $SD_{age} = 11.9$), recruited from Amazon Mechanical Turk[7], took part in Experiment 1. The task took 15 minutes (SD = 8.7) on average and participants were paid at a rate of \$6 an hour. The task interface was programmed in Adobe Flash 5.5.[8] Demos of all three experiments are available at `www.ucl.ac.uk/lagnado-lab/el/oad`.

Stimuli and model predictions. Participants made judgments about nine devices in total. For each device they saw evidence in the form of a single, replay-able video clip. All clips began with a 500 ms interval after which the first component(s) activated. The clip then lasted another 1000 ms whereupon the final component(s) activated. We chose a range of clips in which A occurred at the start and E at the end, varying where B fell in between the two (see Figure 7a, clips 1–7), and then two clips in which E occurred earlier than B (clips 8 and 9). We obtained model predictions by computing the posterior for $p(\mathcal{S}|\,\mathbf{d})$ for Order_{NS} and Order_{S}, and Delay_{P}, assuming learners began each problem with

[7]Mechanical Turk (`http://www.mturk.com/`) is a web based platform for crowd-sourcing short tasks widely used in psychology research. It offers a well validated subject pool, diverse in age and background, suitable for high-level cognition tasks (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Hauser & Schwarz, 2015; Mason & Suri, 2012).

[8]Flash has been shown to be a reliable way of running time-sensitive experiments online (Reimers & Stewart, 2015). We checked the time-accuracy of our code during development finding it highly accurate.

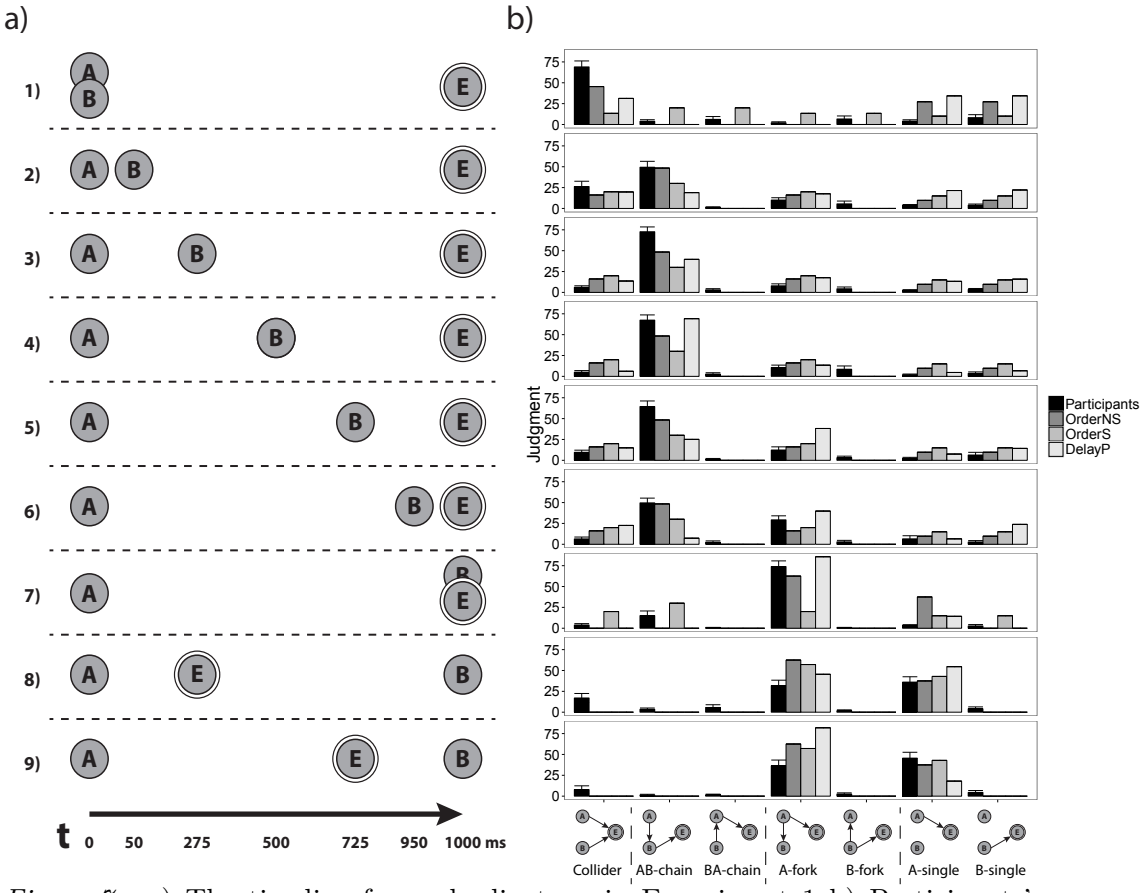

Figure 7. a) The timeline for each clip type in Experiment 1 b) Participants' averaged judgments after viewing each clip (black bars) and predictions by the different models (gray bars). Error bars show standard errors.

a uniform prior across structures and the base rate λ, and a diffuse prior over causal delay parameters α and μ (see Appendix).[9]

Order$_{NS}$ and Order$_S$ model predictions do not vary across clips 2-6 where order was always $A \succ B \succ E$ with both models favoring the AB-chain but Order$_{NS}$ having a stronger preference (see Figure 7b).

They also do not differentiate between clips 8 and 9 (both $A \succ E \succ B$) where both model variants slightly favor the A-fork. Order$_S$ predicts a broad spread across structures for clips 1 and 7, in both cases slightly favoring a chain structure while the Order$_{NS}$ favors the Collider and A-fork, respectively. Sensitivity to timing leads to predictions that differ across clips 2 to 6. Delay$_P$ favors the Collider and A-single and B-single structures when B occurs relatively early, and prefers the chain when B occurs relatively late. Delay$_P$ is also sensitive to the difference in the timing of E between clips 8 and 9, preferring the A-fork if E happens relatively late and the A-single if it occurs early. Finally, it puts more probability mass on the two single structures than the other models.

[9]Note that Delay$_I$ does not make predictions here since it requires repeated evidence to form preferences about the connections.

Table 1
Experiment 1: Order and Delay Models Compared to Participants' Judgments

Model	r	r_s	Mode match	RMSE	N con (N dis)
Baseline				20.1	0
Order_{NS}	0.90	0.76	78%	11.1	12 (10)
Order_{S}	0.71	0.75	56%	16.4	1 (1)
Delay_{P}	0.80	0.64	44%	15.8	3 (4)

Note: Model fits assuming the Collider was conjunctive. r = average Pearson's r correlation between average assignments to structures within each device and model predictions. r_s = average Spearman's rank correlation within problems. Mode match = proportion of problems where participants' modal choice matched model's. RMSE = root mean squared error. N = Number of individuals best correlated by model (con= assuming conjunctive Collider, dis= assuming disjunctive Collider).

Procedure. In the instructions, participants were familiarized with the seven causal structure diagrams, and the response format. Participants then completed the 9 problems in random order. Components A and B were counterbalanced such that on approximately half of the problems faced by each participant their roles were reversed (e.g., B would occur at the start rather than A and their responses flipped for analysis). In each trial, participants observed a single clip of a device and then replayed that same clip. After the fourth replay, participants distributed 100 percentage points across the 7 possible devices displayed at the top of the screen. They were allowed to replay the clip a fifth and final time before finalizing this judgment and moving on to the next device. Participants could only move on if their indicated answers summed to 100%. The causal devices were displayed at the top of the screen in the same order for all problems. For half of the participants, the order of the seven devices was as depicted in Figure 6 while for the other half it was reversed.

Results

There was no effect of counterbalancing on participants' judgments, with no interactions between the A-B counterbalance and participants' assignment of percentage points across the structures, nor with order in which the structures were presented on the screen.

As Figure 7b and Table 1 show, the Order_{NS} model captures participants' judgments best overall here. Comparing participants' responses directly with model predictions, we see that, on average, judgments were well correlated with the Order_{NS}, more so than for Order_{S}, and Delay_{P}. While Delay_{P} beats Order_{S} in terms of Pearson's correlation r, it is a little worse at getting participants' rank order right as shown by the lower Spearman correlation r_s.

As we see in Figure 7b participants assigned some mass to the Collider for clips 8 and 9, suggesting that some participants forgot or disregarded our instruction to think of the Collider as conjunctive (i.e., both causes were needed to generate the effect). To check this we also computed model predictions assuming a disjunctive relationship for the Collider (see Equation A-9 in the Appendix). For the Order models this meant that the Collider likelihood was additionally distributed over patterns 5 and 7. For the Delay models this meant t_E was caused by the earlier-arriving of its two causes. Individually, 12 participants' judgments were closest to Order_{NS} assuming a conjunctive Collider and 10 assuming a disjunctive Collider. Two participants were better fit by Order_{S} and seven by Delay_{P}.

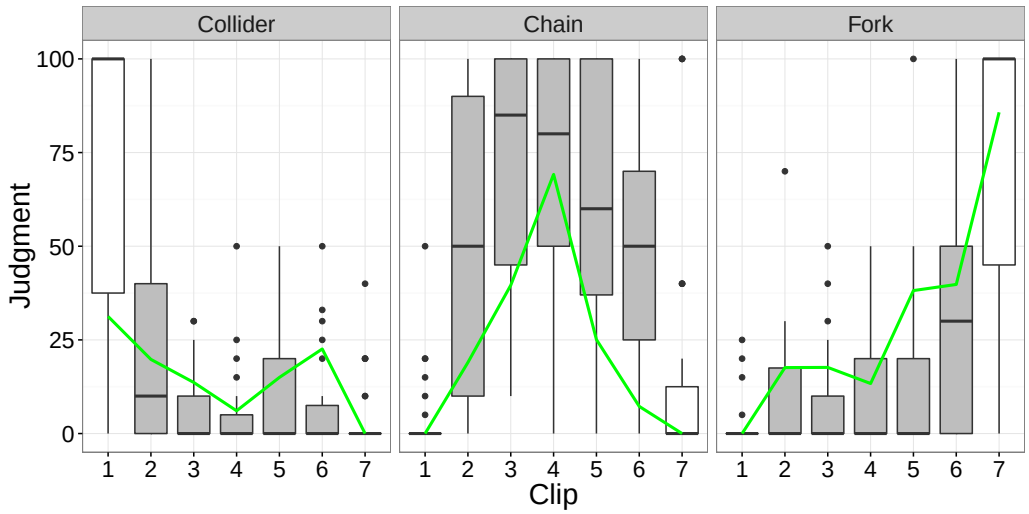

Figure 8. Comparison of probability assignments to Fork, Chain and Collider structures for clips 1–7 (cf. Figure 7), in which B is appears at 0, 50, 275, 500, 725 950 and 1000 ms after A, with E always occurring at 1000 ms. Boxplots show participants' median and upper and lower quartiles, participants with judgments $\pm > 1.5$ interquartile range are plotted separately. Results in text are relative to the six middle bars (gray). Green lines denote Delay_P model predictions.

Overall there was relatively little sensitivity to the exact timing at which B occurred. If we compare patterns 2 to 6 we see that the chain was the modal response across early to late occurrence of B consistent with both Order_{NS} and Order_S predictions. Notwithstanding the dominance of the Order_{NS} model in explaining predictions, there was some evidence of sensitivity to event timings. Figure 8 shows participants' probability assignments to the Collider, chain, and fork for clips 1–7. For clip 2 where B happens right after A, participants assigned some probability to the Collider structure. For clip 6 where B happens right before E, participants assigned probability to the fork. This timing sensitivity is revealed by fitting mixed-effect models to the points assigned to the Collider, the chain, and the fork across clips 2–6, with random means for participants. All three structures' assignments vary across these clips (Collider: $\chi^2(4) = 12.7, p < .013$; AB-chain: $\chi^2(4) = 9.5, p = .05$; A-fork: $\chi^2(4) = 27, p < .0001$) while the order models do not differentiate between these clips. Furthermore, we see hints of the bimodal shape for Collider assignments predicted by Delay_P. For the model this is a consequence of the conjunctive combination function (Equation 6) under which clip 5 is consistent with equal (e.g., 275ms) causal delays $t_{A\to E}$ and $t_{B\to E}$ with A's influence arriving earlier and "waiting" for B's, while not a perfect match to either chain or fork. As expected, chain judgments peaked when t_{AB} and t_{BE} are the same (clip 4) and the fork when t_{AE} and t_{BE} are the same (clip 7).

Discussion

In Experiment 1, we saw that participants' one-shot structure judgments were well explained by a simple model that only uses event order. The model predictions were not perfect though. Order_{NS} underestimated participants' strength of preference for the Col-

lider in clip 1, chain in 3-5 and fork in clip 7, and assigned more weight overall to the A- and B-singles. One possible explanation is that participants might have found some of the structures more or less likely a priori than others. Alternatively, participants might have expected A-single and B-single devices to generate clips in which one of the causal components never occurs even though they were told this would not happen in the instructions. Furthermore, the fact that A and B are perfectly simultaneous in clip 1, might have been seen as evidence for a common causal mechanism — for example some prior mechanism that ensures that the joint causes in the Collider occur in lock-step rather than occurring independently at different times.

The fact that participants' structure preferences were stronger than what was predicted by Order_{NS} might relate to the fact that they replayed each clip several times. Some participants might have treated this as repeated evidence leading to stronger predictions. However, this does not explain the spread of probability in clips 2 and 6.

Participants' judgments shifted over clips 2 to 6 as predicted by Delay_P. This is evidence for some sensitivity to timing, however it was not sufficient to alter many participants' modal judgments away from those predicted by Order_{NS}. Figure 8 shows an inverted U pattern for the chain across clips 2 to 6, rather than the inverted V shaped curve predicted by Delay_P. An explanation for this is that people have limited ability to detect differences between interval lengths, with the modest differences between t_{AB} and t_{BE} in clips 3–5 falling below this threshold. Generally, participants exhibited a robust preference for the chain structure whenever activations occurred sequentially.

In Experiment 1, participants had very little evidence to go on. Having observed a device in action only once, one cannot experience its full range and variability in behavior. Furthermore, single observations limit the scope for forming expectations about delays. In fact, the timings in Experiment 1 were only useful predicated on the Delay_P assumption that all cause–effect relationships between the components within a device have the same means and variances.[10] Thus, to better investigate the adequacy of the Order and Delay models, we now turn to extended learning, where participants observe multiple different clips of the same device and need to integrate the evidence to narrow in on the true causal structure.

Experiment 2: Integrating evidence

In this experiment participants saw several different clips for each causal device. To explore how participants integrated evidence, and to separate the predictions of our two order-based models Order_{NS} and Order_S, we manipulated the order in which components activated during each clip. Participants saw several pieces of evidence, made an initial judgment, and then were able to update their judgment after some additional evidence. This procedure allows us to explore how learners revise their beliefs as they receive more evidence.

We hypothesized that participants' deviations from model predictions in Experiment 1 could be partly due to their having different assumptions about which structures are a priori more likely than others. Another possibility is that while many participants may

[10] Although we note that participants could have formed delay expectations across the task in a hierarchical model fashion.

be relying on temporal order, they might still distribute their likelihood differently than simply dividing it evenly across order-consistent patterns, in particular they might think qualitative patterns that imply reliable delays are more likely than those that do not. We test both of these questions directly in Experiment 2 by eliciting participants' priors and order-dependent likelihoods alongside having them make posterior judgments. This allows us to assess the relationship between prior beliefs, assumptions about the likelihood of different patterns, and posterior inferences on the level of individual participants.

Methods

Participants. Forty participants (19 female, $M_{age} = 30.8$, $SD_{age} = 7.4$) were recruited from Amazon Mechanical Turk as in the previous experiments. The task took 27.0 minutes ($SD = 16.6$) on average and participants were paid at a rate of \$6 an hour.

Stimuli and model predictions. In this experiment, we created evidence sets for 8 different "devices". For each device, participants were presented with four patterns of evidence (see Table 3). They were asked to provide a first judgment after they had seen the first three patterns of evidence, and were then given the chance to update their judgments after having seen the fourth pattern. We selected patterns such that, for five of the devices, our models predicted a strong shift in belief between the first and the second judgment, while for the other three, little or no shift was predicted.

For example, for device 4 (Table 3) participants first saw patterns 1, 2, and 5 ($AB \succ E$, $A \succ B \succ E$ and $B \succ A \succ E$) resulting in a strong prediction by both the Order_{NS} and Order_{S} models that participants will favor the Collider. Finally, participants saw pattern 4 ($A \succ E \succ B$) which is incompatible with the (conjunctive) Collider model, meaning that both models predict a dramatic shift to A-single — the only remaining structure that is consistent with all four patterns (Figure 3 middle row). For three of these five devices the same shift was predicted by both Order_{NS} and Order_{S}, whereas for the other two a different shift was predicted. We only used sets of patterns that did not lead any of the considered models to rule out all the causal structures.

In addition to whether each set of patterns led to a large predicted shift between participants' first and second judgments, we also selected sets of evidence for which the most likely structure differed depending on whether or not participants made the assumption that causes and effects can occur simultaneously. Thus, Order_{NS} and Order_{S} disagreed about the most likely structure for one or both judgments on 2 of the 8 devices (see Figure 3c for an example).

Since we elicit individuals' priors and order-based likelihoods, we can construct an individual order-based model Order_{IV} that makes predictions about $P_{IV}(\mathcal{S}|\,\mathbf{d})$ given the qualitative order of events and each participant's subjective likelihoods $P_{IV}(\mathbf{d}\,|\mathcal{S})$ and prior $P_{IV}(\mathcal{S})$.

In the experiment, we drew intervals between components independently, effectively averaging out any effect of specific timings at the group level. Because each participant experienced different timings, and might have different priors, the Delay_{P} model also makes slightly different predictions for each participant.

Procedure. After reading the instructions, participants had to successfully answer comprehension check questions to proceed. The order in which the devices were presented was randomized between participants. However, the order of clips for each device was

Table 2
Experiment 2: Possible Temporal Order Patterns

Pattern	1	2	3	4	5	6	7
Order	$AB \succ E$	$A \succ B \succ E$	$A \succ BE$	$A \succ E \succ B$	$B \succ A \succ E$	$B \succ AE$	$B \succ E \succ A$

Table 3
Experiment 2: Evidence Sets (1^{st} - 4^{th} Piece of Evidence) for the 8 Devices

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1st	1	2	2	1	2	2	2	2
2nd	2	2	3	2	2	2	2	2
3rd	5	2	4	5	2	2	2	2
4th	2	2	3	4	4	5	1	6
Shift	N	N	N	Y	Y	Y	Y	Y
Different				N	N	N	Y	Y

Note: The numbers in the rows from 1st to 4th refer to the temporal order patterns shown in Table 2. The roles of components A and B were counterbalanced (e.g., pattern 2 $A \succ B \succ E$ becomes pattern 5 $B \succ A \succ E$) and responses re-coded. *Shift* shows whether a change of MAP judgment is predicted by one or both Order models (N)o/(Y)es. *Different* shows whether this shift is predicted to be different between Order_{NS} and Order_{S}.

always as shown in Table 3. We varied the interval between each activation, drawing each from a uniform distribution between 200 and 1200 ms. The clips used in the experiment varied in total length between 1189 and 3094 ms depending on these intervals and whether there were three staggered component activation events (patterns 2, 4, 5 and 7) or only two (patterns 1, 3 and 6). We counterbalanced two presentation orders of the seven structure hypotheses shown at the top of the screen between participants (Figure 6a).

In addition to the *posterior judgment phase* from Experiment 1, we added an initial *prior judgment phase* in which participants were asked to assign 100% points across the seven structures to indicate how probable they thought each of the different structures was a priori (see Figure 9a). In the *posterior judgment phase*, participants made judgments for 8 devices. They were provided with the qualitative visual summary of the clips they had seen. Finally, participants completed a *likelihood judgment phase*. In this phase, participants made seven additional percentage allocations, one for each causal structure. For each allocation, they were shown one of the seven structure diagrams. They were then asked: "Out of 100 tries, how often would you expect this device to activate in each of the following temporal orders?" Participants distributed 100%-points across the different temporal order patterns (see Figure 9b). The order in which participants were asked about each structure, and the order in which the different temporal patterns appeared on each page were randomized between subjects.

When making their posterior judgments, participants were provided with a qualitative summary of the clips they had seen so far (similar to the those in Figure 12a).

Participants were instructed that clicking on the "Start" button constituted the cause of any parentless components in the model. This was indicated in the structure hypotheses by the addition of arrows connecting to any parentless components in each diagram (cf. Figure 9a).

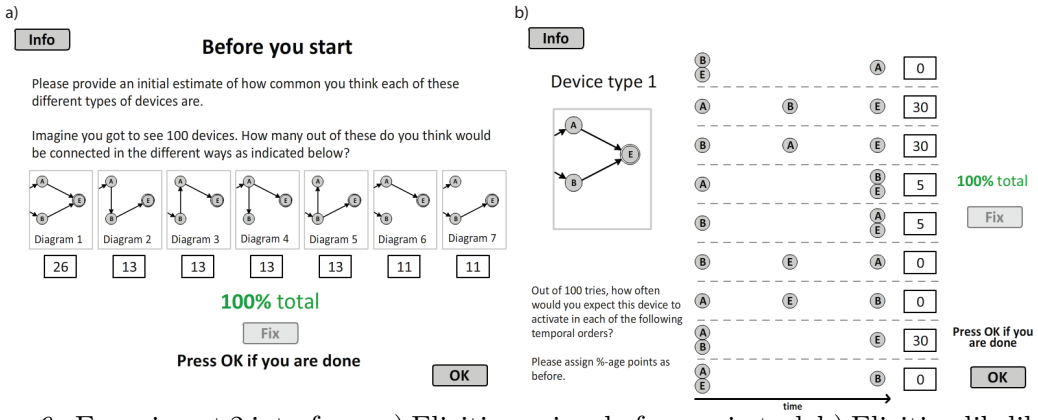

Figure 9. Experiment 2 interface. a) Eliciting priors before main task b) Eliciting likelihoods after main task.

Results

We will discuss the results from the *prior judgment phase*, *likelihood judgment phase*, and *posterior judgment phase* in turn.

Prior judgment phase. 21 of the forty participants' priors differed significantly from a uniform according to χ^2 tests, with Bonferroni corrected significance (i.e., $p < \frac{.05}{40}$). After removing two participants who assigned 0% to more than half of the structures, we performed a cluster analysis on the remaining 38 participants, finding three clusters.[11] Twenty-two participants assigned roughly equal weight to all seven options (see Figure 10). Twelve assigned approximately double to the Collider compared to the rest of the structures. Four other participants formed a third cluster with no apparent systematicity in their priors.

The 12 participants who gave more mass to the Collider structure might have been thinking in terms of *types* of structure, dividing evenly across Colliders, chains, forks and single, then subdividing within each type. This could explain their putting more prior weight on the Collider, since it is the only structure within its class. By splitting the resulting probabilities across class members, the Collider ends up with greater prior probability due to being a unique member of its class.

Table 4
Experiment 2: Likelihood Judgment Model Fits

Model	r	r_s	Mode match	RMSE	N
Baseline			11%	15.4	1
Order_{NS}	0.92	0.78	71%	8.9	11
Order_{S}	0.57	0.80	43%	12.8	4
Delay_{P}	0.98	0.81	100%	7.3	24

Note: r = average Pearson's r correlation between average assignments to structures within each device and model predictions. r_s = average Spearman's rank correlation within problems. Mode match = proportion of problems where participants' modal choice matched model's. RMSE = root mean squared error. N = Number of individuals best correlated by model.

[11] This was established by fitting a Gaussian finite mixture model using R's `mclust` package.

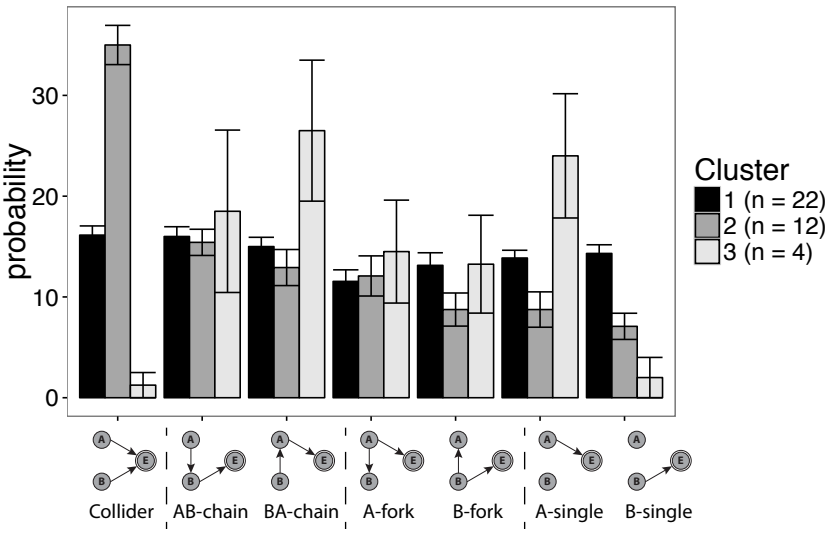

Figure 10. Elicited priors split into clusters as detailed in text. Error bars show standard errors.

Likelihood judgment phase. Likelihood judgments were most highly correlated with marginal likelihoods of the patterns under Delay_P ($r = .98$), followed by Order_{NS} with Order_S considerably lower (see Table 4).[12] Inspecting Figure 11, reveals that the Delay_P based likelihoods captured the fact that participants assign more probability to the patterns implying reliable delays (more to pattern 1 than patterns 2 or 5 for the Collider, and more to pattern 3 than 2 or 4 for the A-fork, and similarly for the B fork).

To check whether participants largely made the same assumptions about the devices as our models, we checked how frequently they assigned likelihoods to patterns ruled out under all of the models we consider. Overall, participants assigned much less likelihood to these patterns (10.5%) compared to the 44% expected from random allocation. However, eighteen participants assigned some likelihood to patterns ruled out by Order_{NS}, Order_S and Delay_P, assigning an average of $7.4 \pm 13\%$ of their points to 5.3 ± 10 of the 24 patterns. Of these, the most frequently were $A \succ E \succ B$ and $B \succ E \succ A$ under the Collider, with nonzero likelihoods assigned by 12 and 14 participants respectively. As a result there was a higher probability of assigning non-zero likelihoods to patterns ruled out by our models under the conjunctive Collider than on average over the other structures $\chi^2(1) = 5.1, p = 0.02$. This confirms our suspicion that some participants did not make the conjunctive assumption when reasoning about the Collider, in spite of instructions.

Posterior judgment phase. We analyzed the posterior judgments by comparing linear mixed models with random intercepts for participants, and structures within participants. By design, neither device (1:8) nor judgment (1^{st} vs. 2^{nd}) can have a main effect on assignment of % points. This is because judgments were constrained to add up to 100% across the structures. Instead, effects are indicated by interactions between these

[12] We assumed the same parametrization as in Experiment 1, and encoded the timings implied by the depictions of the order patterns (e.g., Figure 11a) assuming they represented a total interval of 1400 ms, with 700 ms between the initial and middle events for patterns 2, 4, 5 and 7, corresponding to the mean interval between events in the task.

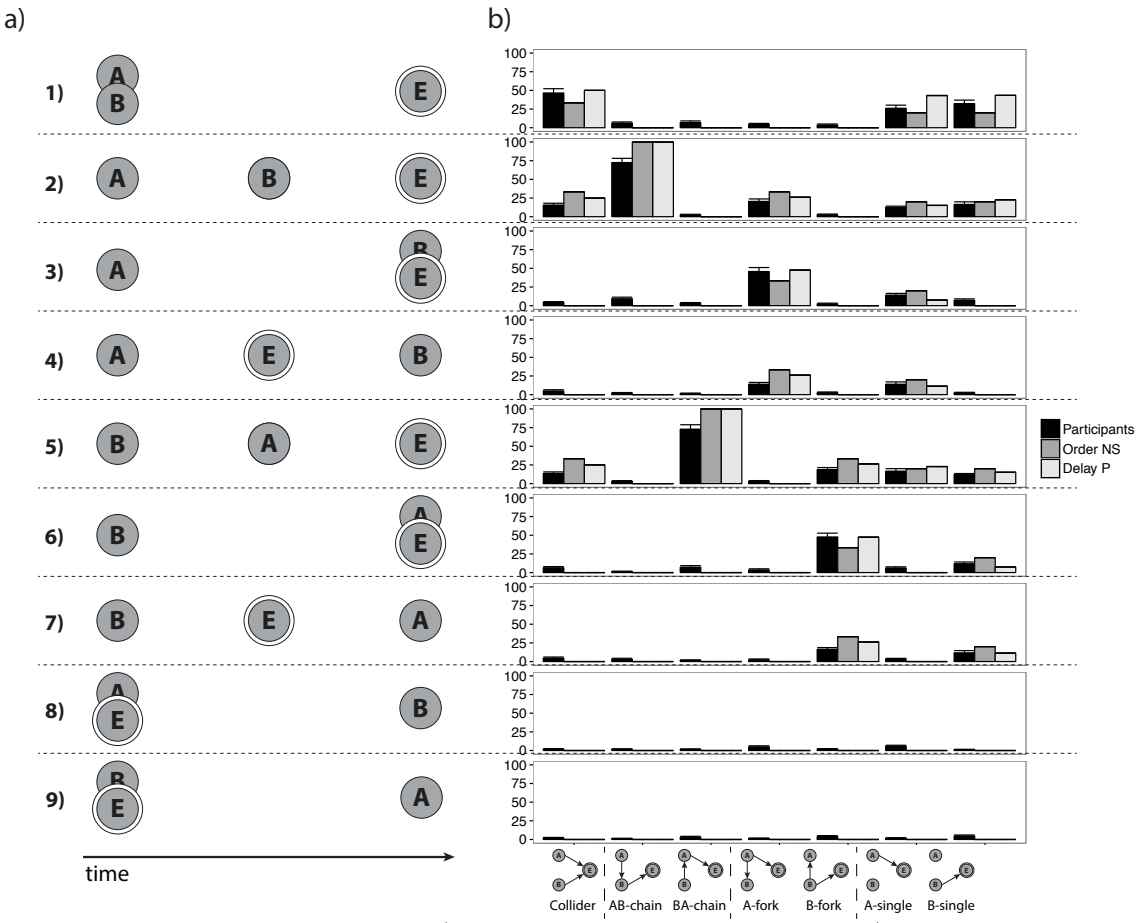

Figure 11. Elicited likelihoods. a) Nine temporal order patterns b) Participants elicited likelihoods compared with those of Order_{NS} and a variant or Order_{NS} type that distributes likelihoods across *types* of consistent patterns — lumping together $A \succ B \succ E$ and $B \succ A \succ E$ for the Collider, and $A \succ B \succ E$ and $A \succ E \succ B$ for the forks. Error bars show standard errors.

different factors and the assignments across the structures. Structure interacted with device $\chi^2(55) = 1384, p < .0001$, confirming that judgments were affected by the different evidence sets. Judgment (1^{st} versus 2^{nd}) also interacted with device $\chi^2(7) = 99, p < .0001$, and there was a three-way interaction between judgment, structure and device $\chi^2(49) = 286, p < .0001$ confirming that the impact of the final piece of evidence was different for some devices than others. The complexity of these interactions prohibits direct interpretation but we can compare judgments' to the predictions of Bayesian updating based on participants' elicited priors and likelihoods, either with or without additional sensitivity to the intervals.

Participants' average posteriors were very closely correlated with the predicted average over posteriors based on the priors and order-based likelihoods they provided (Order_{IV}). By computing these posteriors then averaging over participants, we get a $r = .95$ correlation with judgments and a *RMSE* of 7.0% compared to baseline of 14.3%. It does not make sense to average the Delay model posteriors in this experiment since timings dif-

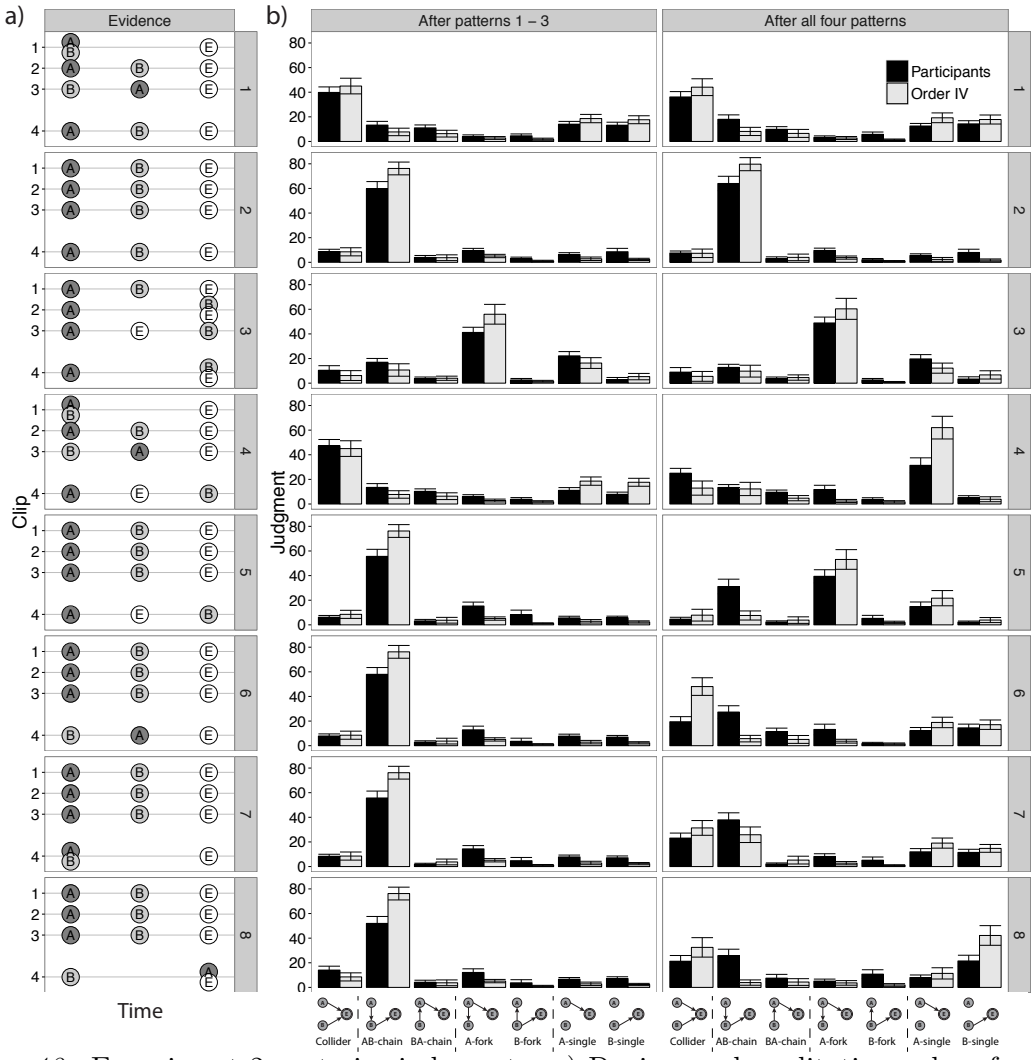

Figure 12. Experiment 2 posterior judgments. a) Devices and qualitative order of activations for each. *Note*: Exact timings were drawn at random from Uniform$(200, 1200)$ for each participant in this experiment and so are not shown in full. b) Participants' posterior judgments (black bars) compared to a model based on individually elicited priors and order-based likelihoods Order_{IV} (gray bars). Left hand column, judgments after viewing 3 clips, right hand column judgments after all four clips. The Order_{IV} bars omit cases in which participants' chosen likelihoods and priors led to all hypotheses being ruled out. Error bars show standard errors.

fered between participants. However, we can check for timing sensitivity at the level of individuals. Here, we find that the most participants' posteriors are still best described by the Order_{IV} model that combines the priors and order-based likelihoods they provided themselves (28/40).[13] However 10 were better described by the posteriors under Delay_P

[13]For 14% of first and 27% of second judgments, all structure hypotheses were ruled out based on combining individuals' priors and likelihoods. This happened at least once for 19 out of the 40 participants. To allow

suggesting some additional sensitivity to experienced timing.

Inspecting Figure 12 we see that aggregated participant posterior judgments are typically a little less peaked than the aggregated Bayesian posterior predictions, even though these were based on their priors and likelihoods, and account for the heterogeneity of assumptions people made about the task. In particular, where we chose a fourth clip such that we predicted a modal shift between the first and second judgments (devices 5-8), we see considerable residual percentage points for the previously favored structure to that which the models favor after. For example, for device 5, participants' priors and likelihoods suggest they should strongly favor the A-fork after viewing the final clip, but participants move only around half the probability mass, leaving a considerable amount "behind" on the previously favored AB-chain, which all our models consider to be ruled out. This suggests that participants were generally somewhat conservative in their updates. Their beliefs were moved less by the evidence than their priors and likelihoods would suggest they should be (Edwards, 1968). To test this more thoroughly we considered a variant of Order_{IV} that updates its beliefs conservatively.

We can model conservatism within the Bayesian framework through addition of unbiased noise to participants' likelihood functions, such that patterns that should be ruled out by a structure given one's assumptions, instead retain some ϵ probability. If participants are generally conservative, we expect such a model that incorporates noise into the likelihoods for each observation to better explain their final judgments.

To do this, we created noisy likelihoods by mixing each participant's reported likelihood function with uniform likelihoods (with $\frac{1}{9}$ over the 9 patterns of data participants distributed over for each structure) to a degree controlled by a free parameter $\epsilon \in [0, 1]$ (i.e $P_{IV}(\mathbf{d}\,|s)^{cons} = (1-\epsilon)P_{IV}(\mathbf{d}\,|s) + \frac{1}{9}\epsilon$).[14] We fit ϵ to each participant's data separately, by maximizing the correlation between the prediction given Bayesian integration of the priors and likelihoods they reported, and their own posterior judgments.[15] We found that 32/40 participants had a non-negligible best-fitting ϵ parameter (> 0.01), indicating conservatism in their evidence integration relative to the Bayesian ideal. The mean ϵ was .36 ($SD = .35$). Inclusion of conservatism increased the aggregate model correlation of Order_{IV}^{cons} to $r = .97$ $RMSE = 6.5\%$ compared to Order_{IV}'s $r = .95$, $RMSE = 7.0\%$.

Discussion

In Experiment 2, we attained a clearer picture of the sources of variability in people's causal structure inferences. Many participants reported priors that distributed probability mass uniformly at the level of *types* of structures rather than response options. The Collider was the only unique structure (since there were two chains, two forks and two singles), and it was judged to be a priori more likely than the rest of the structures by many participants. This suggests that these participants generated uniform prior probabilities based on more abstract representations of the causal structures and evidence patterns, rather than taking the option set we provided as distributionally representative.

comparison we simply had the Order model predict a uniform distribution over the hypotheses in these cases, guaranteeing a correlation of 0 for that device — the same as the Baseline model.

[14] For $\epsilon = 1$, $P_{IV}^{cons}(\mathbf{d}\,|\mathcal{S})$ is uniform and therefore results in no belief change.

[15] We use the Brent (2013) algorithm to do the optimization as implemented by R's `optim` function

Participants found structures that exhibited equal delays more likely than unequal delays. We were able to capture this very well by our Delay_P model which favors structures that imply causal delays that are more similar on average (with a $r = .98$ correlation with the aggregate patterns and a better fit than the other models we considered for 24/40 participants). Participants' made these likelihood judgments after having completed the posterior judgment phase. It is thus possible that they tried to make their likelihood judgments consistent with the posterior judgments they had provided in the previous phase of the experiment.

Interestingly, despite distributing likelihoods in a way that suggested they preferred equal delays across devices' components, participants still appeared quite insensitive to exact event timings. The majority of participants' posterior judgments were better described by Order_{IV} than Delay_P suggesting that participants paid little attention to exactly how far apart in time the events were in the clips. We note here though that the design of the experiment might have nudged people toward this behavior. We provided summaries showing the qualitative order of events in Experiment 2 while the exact event timings were only represented in the clips themselves. This which may have encouraged participants to focus predominantly on order. Furthermore, by selecting clips that provided lots of order information, the resulting data was not distributionally representative of reliable generative gamma delays.

We found that we can capture participants judgments even better by positing that they were somewhat conservative in their integration of the evidence they observed, over and above what was implied by the likelihoods they provided. Conservatism relative to Bayesian predictions is a consistent psychological finding (Bramley, Lagnado, & Speekenbrink, 2015; Edwards, 1968; Fischhoff & Beyth-Marom, 1983). In this task, it could reflect a number of things. Participants may have suspected that the devices might change structure over time, and so not want to rule out a possibility that could later be true. They might also distrust what they were told in the instructions, have forgotten or be unsure about them (Corner, Harris, & Hahn, 2010). Under-updating of judgments might more fundamentally be a a consequence of their processing limitations, either directly, or as a way of compensating for the possibility of having made perceptual or memory errors about the evidence they had seen.

Our qualitative order models did well in explaining participants' inferences in the tasks we have looked at so far, even explaining evidence integration over multiple trials where there is, in principle, enough timing evidence to start to form expectations about the delays. However the experiment emphasized order information by using non-representative delays and providing qualitative visual summaries of the evidence during posterior judgments, yet Delay_P still outperformed Order_{NS} for some participants. Finally, the close correspondence between Delay_P and participants' qualitative likelihood ratings clearly show that timing matters, even if their role here was predominantly limited to shaping peoples' order expectations.

To look more closely at the role of timing, we now turn our focus to a situation where order is non-diagnostic and the only available information comes from the variability and correlation in event timing. This will allow us to assess the extent to which people are capable of using timing information at all, and the adequacy of our normative model in capturing the ways in which people use temporal information.

Experiment 3 — Learning from timing variability alone

In this experiment, we focus on causal inference from timing variability alone. To isolate timing from order cues, we chose a more constrained situation than before, with only two possible structures (an $S \rightarrow A \rightarrow B$ chain and an $A \leftarrow S \rightarrow B$ fork) and evidence where the order of activation of three components was (almost) always the same ($S \succ A \succ B$). We systematically varied the mean and variability of the inter-event timings such that they were more consistent with having been generated by either a chain or a fork under the Delay$_I$ assumption as we describe below.

We hypothesized that participants would be sensitive to these differences and able to use them to distinguish between the two candidate structures. However, we also expected based on the results from the previous experiments, that participants would have a general preference for the chain. While the chain can only produce the $S \succ A \succ B$ pattern, the fork is more flexible. We also hypothesized that participants would find it more difficult to draw inferences from quantitative differences in time intervals, versus the more obvious and definitive qualitative differences in event order. Thus, we predicted that participants would be more uncertain overall in their posterior judgments. To assess how well participants detect and track timing variability across tests and hypotheses, we first elicited judgments based on simply experiencing the timings. Afterwards, we provided participants with summaries of the trials detailing all the timings visually, and allowed them to update their judgments. The idea was that providing participants with summaries, would eliminate any potential memory effects, or effects resulting from perceptual noise associated with encoding the timings, providing a helpful comparison to the judgments based on experience alone. Generally, we expected that participants' preference for one of the two structures to become stronger and more normative after having seen the summary.

A further question is whether participants who are able to learn the true causal model are also able to learn the causal delays, such that they can make predictive judgments about what patterns of evidence the device is likely to produce in future tests. To explore this question, the experiment included an additional task where participants had to make a predictive judgment.

Methods

Participants and materials. 104 University College London undergraduates (87 female, $M_{age} = 18.8$, $SD_{age} = 0.81$) took part in this experiment under laboratory conditions as part of a course requirement. The task took 23.0 minutes ($SD = 3.1$).

Stimuli. Participants had to judge whether a device was a $S \rightarrow A \rightarrow B$ chain or a $A \leftarrow S \rightarrow B$ fork. Both chain and fork structures shared an $S \rightarrow A$ connection, but differed in whether they had an $S \rightarrow B$ or an $A \rightarrow B$ connection. This implies that t_B could be explained by one of two delay distributions: either $t_{S \rightarrow B}$ or $t_{A \rightarrow B}$. Under the independent Delay$_I$ model, this results in a preference for one of the two structures, depending on which of these inferred delay distributions can assign more likelihood to the evidence (marginalizing over its unknown parameters).

In order to construct the evidence, we first created two generative chain (*chain1* and *chain2*) and fork devices (*fork1* and *fork2*) by augmenting each connection with a delay distribution (see Figure 13a). All four devices shared an $S \rightarrow A$ connection with

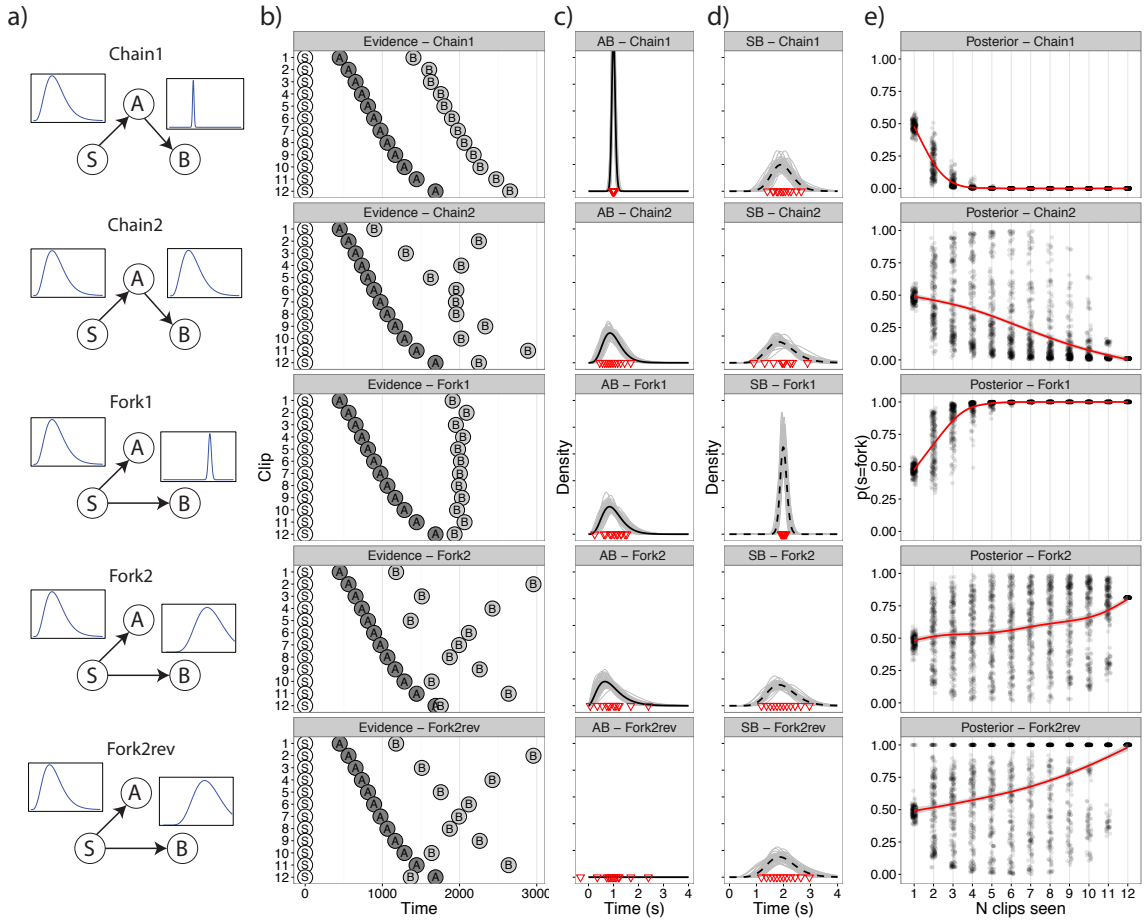

Figure 13. Experiment 3 stimuli and model predictions. a) Graphical representation of the five device types. b) Plot showing the 12 patterns generated for each device. c) Red inverted triangles: t_{AB} for patterns 1:12. Gray lines: $P(G_{A\to B}|\,\mathbf{d})$ for a posterior sample of αs and βs. Dashed black line: The posterior marginal likelihood of $G_{A\to B}$. d) As in c) but for $G_{S\to B}$ under the fork structure. e) Posteriors $P(s = fork|\,\mathbf{d})$ for progressively more evidence. Individual dots for the samples of evidence seen by participants, lines smoothed average (using the general linear additive model with integrated smoothness estimation `gam` from R's `mgcv` library). *Note*: Individual points are jittered along the x-axis to increase visibility.

delay distribution $G_{S\to A}(\alpha = 5, \mu = 1000\text{ms})$. Concretely this meant that A would occur an average of 1000ms seconds after S but with considerable variability. We then chose distributions for $A \to B$ for the chains and $S \to B$ for the forks such that interval between S and B t_{SB} was 2000ms on average, but the shape and extent of the variability in the timing of B depended on the underlying connections.

In *chain1* there was a near-constant $G_{A\to B}(\alpha = 1000, \mu = 1000\text{ms})$, while in *chain2* $t_{A\to B}$ had as much variability as $t_{S\to A}$. In *fork1*, the $S \to B$ connection had a near-constant 2 second delay $G_{S\to B}(\alpha = 1000, \mu = 2000\text{ms})$ while in *fork2* the delay was variable $G_{S\to B}(\alpha = 10, \mu = 2000\text{ms})$.

We used these four generative devices to select sets of 12 clips used as evidence. To ensure that the selection of clips was representative for the generating distributions, we took 12 equally spaced quantiles from each distribution.

To ensure that the delay draws for $G_{S\to B}$ (or $G_{A\to B}$ for the forks) were independent of those for $G_{S\to A}$, they were paired in counterbalanced order. The resulting sets of evidence are depicted in ascending order of t_{SA} in Figure 13b. Finally, we included a variant of *fork2*, named *fork2rev*, which included a single order reversal trial. This allows us to compare the respective strengths of order and timing cues.

Model predictions. We used Delay_I to obtain a posterior joint distribution over the true structure (i.e., fork or chain) and its associated parameters.[16] We obtained posterior predictions by averaging over the parameters. These predictions are normative in the sense that the Delay_I model inverts the true generative model. Figure 13e shows how these predictions change with each additional clip seen. Because we randomized the order of the clips, there is variability in what evidence the model has received so far. Each point in the plots shows the predicted posterior given the evidence an individual participant has seen up to this point. The red line shows the averaged predicted posterior. By the 12th clip, all participants have seen the same evidence so the predictions converge.[17]

Figure 13e shows that the model rapidly infers that the true model is a chain for *chain1* and a fork for *fork1*. Looking at the predictive distribution subplots (Figure 13c and d), we see that this is due to the model's ability to fit a tighter distribution onto the experienced timings under the true model, assigning less mass to all the data points while they are more spread out and unevenly distributed under the alternative structure. Under the noisier *chain2* and *fork2* evidence, the model forms the correct preference but does so much more slowly, retaining significant uncertainty even after 12 clips for *fork2*, where the delay distribution is only slightly less variable under the fork structure than the chain. Finally, for *fork2rev* the predictions are the same as *fork2* until the order reversal trial is seen and the chain is ruled out. This becomes increasingly likely on later trials and certain after all 12 clips. Thus normatively, we expect more points to be assigned to the chain structure for *chain1* and *chain2*, than for *fork1*, *fork2* and *fork2rev*; more to *chain1* than the more difficult to infer *chain2*. Likewise, we expect more points to be given to the fork structure for *fork1* than *fork2*. Finally, since the order cue in *fork2rev* rules out the chain we expect judgments here to be more strongly in favor of the fork structure than for the other fork patterns.

Procedure. Participants were instructed about the two possible causal models, the interface, the number of problems they would face, the number of tests they would perform for each problem, the presence of delay variability, and the independence of variability between different connections. Participants initiated the system by clicking on the "S" component and watching when the other two components activated (see Figure 14a). To familiarize participants with the delay variability, they interacted with four two-component devices during the instructions, each with a single cause and a single effect. They tested

[16] We used the Delay_I variant of our delay model because the Delay_P variant assumes that all delays share the same parameters, and participants were explicitly instructed that this was not the case.

[17] We used MCMC to estimate these posteriors without specifying any prior on delay parameters. In the appendix we compare these to Simple Monte Carlo sampling predictions under a variety of priors. This allows us to assess the impact of prior choice in Experiments 1 to 3.

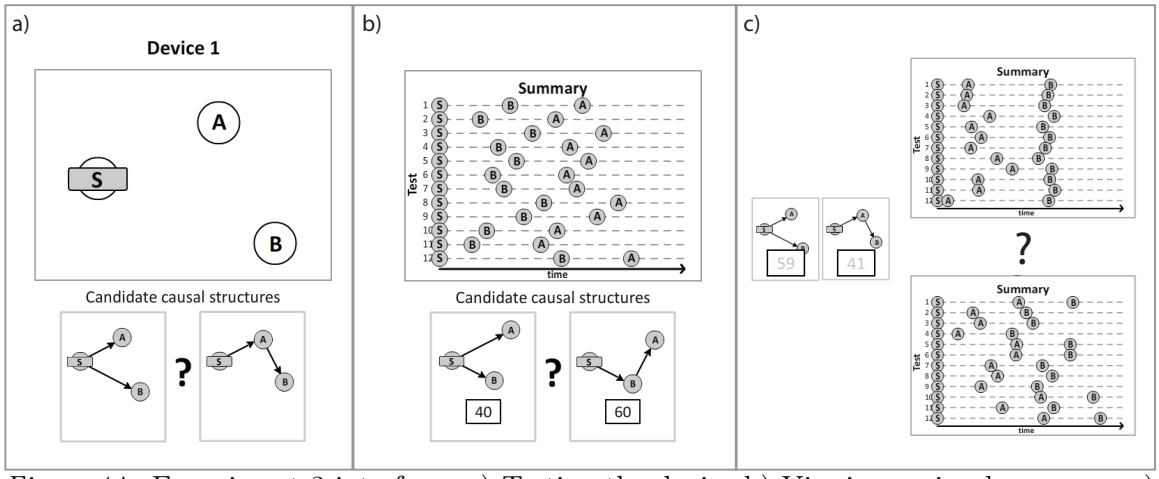

Figure 14. Experiment 3 interface. a) Testing the device b) Viewing a visual summary c) Making a predictive judgment.

each device at least 4 times. There were two pairs with short ($\mu = 1s$) delays, one near-constant and one variable, and two with longer ($\mu = 2s$) delays, likewise one near constant and one variable. Participants were also instructed that the variability of the delays of the different components of a device were independent such that an unusually long $t_{S\to A}$ would not imply that there would be an unusually long $t_{S\to B}$ or $t_{A\to B}$. Before proceeding to the main task, participants had to correctly answer comprehension check questions.

All participants faced each of the 5 problem types twice, once as detailed in Figure 13 and once in with the labels and locations of A and B reversed (as in Figure 14b). Thus, there were 10 within-subjects test problems overall. On each test problem, participants watched 12 clips in a random order. For each problem they made 3 causal judgments. They made their first judgment after the 6th clip, their second after all 12 clips, and a final judgment after seeing a visual summary of the timelines of the clips they had seen (similar to the quantitative summary in Experiment 2, see Figure 14b). Participants gave their causal judgments by distributing 100% points across the two structures. During trials 7–12, participants' initial response remained visible but grayed out in the response boxes. They then had to interact with one of the response boxes (changing the value or just pushing enter) to unlock the "Continue" button on the second and third judgments.

In addition to eliciting structure judgments for 10 problems within subjects, we also elicited predictive judgments on one additional problem which was varied between subjects. On this final problem, participants either saw evidence from *chain1* or *fork1*, in a new order. We selected which evidence was seen at random between subjects (45 out of 104 subjects saw *chain1*, the rest saw *fork1*). The first and second judgments were identical to the previous problems, but instead of seeing the visual summary, participants were presented with two side-by-side visual summaries of new draws of 12 patterns, one generated by a *chain1* structure and one by a *fork1* structure (See Figure 14c). They were then asked to distribute 100% between the two sets of evidence indicating which evidence was more likely to be produced by the current device.

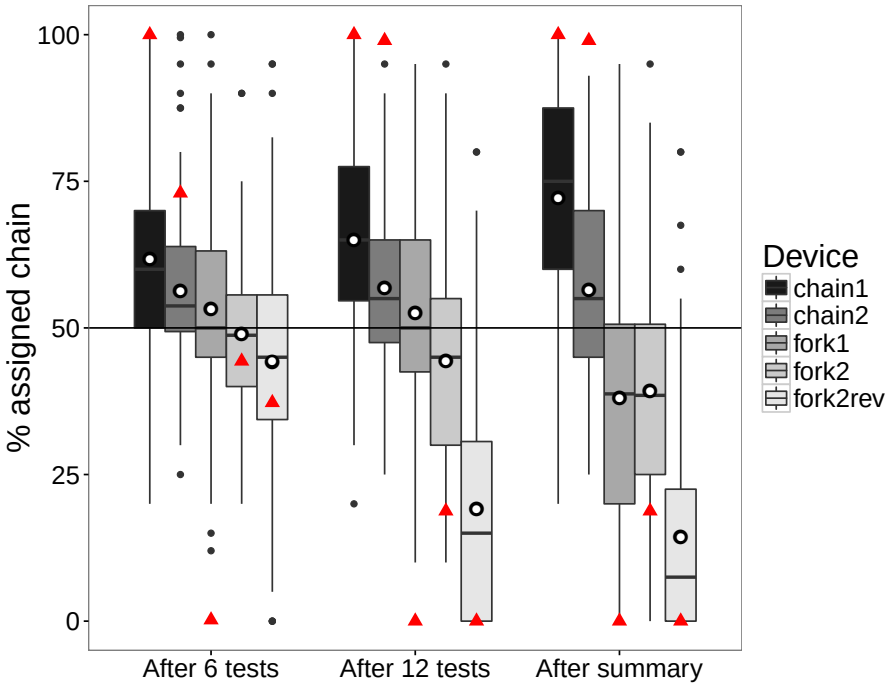

Figure 15. Judgments for the different devices. Boxplots show participants' median and upper and lower quartiles, participants with judgments ± > 1.5 interquartile range are plotted separately. White filled black circles = participant means. Red triangles = Delay_I posteriors.

Table 5
Experiment 3: Main Effects and Planned Comparisons for First, Second and Third Responses

	Response 1	Response 2	Response 3
Main effect (LR)	86***	334***	378***
Planned contrasts			
Intercept	$52 \pm .9\%^{***}$	$46 \pm .9\%^{***}$	$42 \pm .9\%^{***}$
1. Chains vs. Forks	$10.2 \pm 1.3\%^{***}$	$22.2 \pm 1.3\%^{***}$	$33.8 \pm 1.6\%^{***}$
2. Chain1 vs Chain2	$2.7 \pm 1.0\%^{**}$	$4.1 \pm 1.1\%^{**}$	$7.8 \pm 1.3\%^{***}$
3. Fork1 vs Fork2	$2.1 \pm 1.0\%^{*}$	$4.0 \pm 1.1\%^{**}$	$-.6 \pm 1.3\%$
4. Forks1&2 vs Fork2rev	$4.5 \pm 1.1\%^{**}$	$19.6 \pm 1.3\%^{***}$	$16 \pm 1.5\%^{***}$

Note: For main effects we report the likelihood ratio for a model with device type as predictor relative to a model with just an intercept. For each planned comparison we report the size of the effect (%) ± standard error, and level of significance: $^{*} = p < .05$, $^{**} = p < .01$, $^{***} = p < .001$.

Results

Structure judgments. In order to analyze participants' judgments, we must account for the fact that each participant faced each device twice. To do this, we fit linear mixed effects models to all judgments, with participant and device as random effects. To test our specific hypotheses about the differences between devices, we constructed four

orthogonal contrast codes.

These compared: 1. [*chain1*, *chain2*] to [*fork1*, *fork2* and *fork2rev*], 2. *chain1* vs *chain2*, 3. *fork1* to *fork2* and 4. [*fork1*, *fork2*] to *fork2rev*, matching the normative hypotheses described above. The four regressions are summarized in Table 5. All three judgments differed by device type, with the size of these differences increasing for the judgments made after performing 12 compared to 6 tests, and after seeing the visual summary relative to before. For instance, participants assigned 10.2% more percentage points to to the chain diagram when the true structure was a chain, after 6 tests, increasing to 22.2% after 12 tests and to 33.8% after viewing a visual summary of the evidence. On all three judgments, participants assigned significantly more points to the chain diagram for chains than forks. They also assigned more chain points to the reliable than the unreliable chain, and more to the forks that did not exhibit the order cue (*fork1* and *fork2*) compared to *fork2rev*. However, on judgments after 6 and 12 tests, participants assigned *more* points to the chain diagram (i.e., fewer to the fork) for the theoretically easier and "reliable" *fork1* than the "unreliable" *fork2*. After the visual summary, the fork diagram was equally favored for each of these two devices.

Looking closely at the evidence we generated, we see that the difference between *chain2* and *fork2* is very subtle. While the t_{AB} interval is more variable under *fork2* than *chain2* (Figure 13c 2$^{\text{nd}}$ vs 4$^{\text{th}}$ row), t_{SB} is actually also slightly more variable under *fork2* than *chain2* (Figure 13d). Thus, if participants focused only on t_{SB} we would expect them to favor the chain structure for this problem. The fact that participants still form a preference for the chain for *chain2* and the fork for *fork2* based on this subtle difference in t_{AB}, while failing to note the reliable t_{SB} in *fork1*, is suggestive that participants were particularly tuned to monitoring the successive intervals rather than the overall interval. We examine this idea in more detail in the General Discussion in the Section *Sensitivity to timing: Toward a process model*.

Predictions. For the final problem, participants had to predict which of two evidence sets was more likely to be generated by the device they had just learned about. Here, participants favored the chain evidence marginally more when the true structure was *chain1* compared to when it was *fork1* $t(102) = 1.7, p = .044$ (one-tailed). Participants assigned significantly more than 50% to the chain evidence when the true structure was the chain $t(44) = 1.9, p = .029$ (one-tailed) but were not significantly more likely to favor the fork evidence for the fork device $t(58) = -.26, p = .36$ (one-tailed). However, participants' strength of judgment toward the chain (/fork) was not statistically related to their preference for the evidence actually generated by the chain (/fork) $F(1, 102) = .6, r = .08, p = .4$. For example, in Figure 14c the fact that this particular participant assigned 59% to the fork does not mean they will assign more predictive probability to the future fork-generated evidence (top) over the chain-generated evidence (bottom). We note, though, that we did not test participants' predictive knowledge very thoroughly in this experiment. Only a single predictive trial was included, varied between subjects, and there was no incentive or instruction for participants that they should try to learn to predict the devices. Further research is needed gauge the extent to which people can learn to predict the temporal dynamics of causal systems.

In sum, we found that people were able to distinguish between direct and indirect causation (i.e., a fork and a chain) based on the variability and correlation in event timings

alone. However, people found this inference much tougher than making judgments based on having observed different temporal orders of events. In this experiment, some participants reported relatively weak preferences despite having seen considerably more data, and having fewer structure hypotheses to evaluate than in Experiments 1 and 2.

Discussion: Sensitivity to timing

Our Bayesian Delay$_I$ model broadly captured aggregate judgments (see Figure 15). However, there is some evidence that participants may have solved the task in a more heuristic way. Firstly, participants' judgments were much less strong than the normative models' preferences. Secondly, participants had trouble predicting future evidence, suggesting they did not finish each problem with clear expectations about the device's delays. Third, the Delay$_I$ strongly favored the fork structure after seeing only a few clips from *fork1*, while participants remained at chance for this problem until the summary.

Ideal probabilistic structure inference involves maintaining a probability distribution over all candidate hypotheses. This is infeasible in the general case as there is a near-infinite number of possible models. There have been several recent proposals that people maintain a single candidate causal model at a time, stochastically switching when their current model proves strongly incompatible with evidence (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Bramley, Griffiths, Dayan, & Lagnado, in revision). Additionally, Lagnado and Sloman (2006) propose that people often take event order as an initial proxy for causal order. In this section we consider several heuristics based on the idea that participants in Experiment 3 used simpler statistics to identify the generative model without computing the predictions under both structures at once.

Does A predict B?. In general, if A causes B, we expect that the time at which we observe A (relative to its cause S) to be predictive for when we will later observe B (also relative to S). Thus, a reasonable proxy for computing the full posterior is to try and estimate the strength of this predictive signal. In the current context this comes down to a correlation between t_{SA} and t_{SB}, hereafter $\text{cor}(t_{SA}, t_{SB})$. If $\text{cor}(t_{SA}, t_{SB})$ is positive, this is a sign that S's causing of B may be mediated via A — that is, observing an unusually early/late A is a noisy predictor of an early/late B (see Figure 13a). Conversely, if t_{SA} is statistically independent of t_{AB} this is more consistent with the idea that B is caused directly by S as in a fork structure.

Variance under a single structure. Computing a correlation between t_{SA} and t_{SB} across clips might still make too strong demands on perception and storage to be estimated online while watching the clips. The issue here is that the correlation depends on encoding two overlapping intervals for each test, storing them, and comparing their relationship across multiple trials. It is well-established that there are strong limitations on explicit attention and short-term memory which may prohibit such explicit multitasking (Baddeley, 1992; Lavie, 2005). Rather, it seems plausible that learners might only monitor the timings in the clips under a single hypothesis at a time, for example either focusing on t_{AB} if they are currently entertaining the *chain* structure, or t_{SB} if currently entertaining the *fork* structure.

Accordingly, a simpler strategy than comparing models would be to monitor the variance assuming that one or the other structure is true. If this variance seems "too high"

one can reject the structure hypothesis and start monitoring the delays under the alternative structure.

Assuming that participants tend to perceive event order as causal order by default (Lagnado & Sloman, 2006), it is possible that participants found it more natural to monitor $\sigma^2(t_{SA})$ and then $\sigma^2(t_{AB})$ than to monitor $\sigma^2(t_{SB})$ (while ignoring the intervening event at A). Thus, $\sigma^2(t_{SB})$ may effectively have been masked by participants' default tendency to perceive succeeding events as a chain, and thus only encode the delays between directly succeeding events.

Online approximation. Estimating variance of the delays across trials may already be challenging. As we mentioned in the introduction, many models of sequential estimation avoid storing all the data, replacing it with an operation over all the evidence with a simpler adjustment that can be performed as evidence comes in (Halford, Wilson, & Phillips, 1998; Hogarth & Einhorn, 1992; Petrov & Anderson, 2005). We propose a simple model based on this idea here. Average pairwise difference (APD) simply stores the difference between the interval in the latest clip t^k_{XY} and the one before t^{k-1}_{XY}, summing this up across trials. When variance is high this will tend to be high too but it is also sensitive to the order in which evidence is observed, being larger when intervals fluctuate more between adjacent tests.

Each of these measures — $cor(t_{SA}, t_{SB})$, $\sigma(t_{SA})$, $\sigma(t_{SB})$, $\sigma(t_{AB})$, $\text{APD}(t_{SA})$, $\text{APD}(t_{SB})$, and $\text{APD}(t_{AB})$ — assigns a value to the evidence seen at each time point by each participant. Thus, all the measures make different predictions for each participant on the first judgment because the clips seen so far differ between participants. Additionally, the APD measure is computed sequentially and thus creates order effects and results in different predictions for different participants for the second and third judgments, too.

We used all these measures as predictors of the number of percentage points assigned to the chain structure on each judgment with a prediction of zero indicating 50% chain (50% fork). This means that measures which support the chain have positive weights and measures that support the fork have negative weights, and the intercept indicates a baseline preference for one or the other model.

We hypothesized that one or a combination of these simpler measures $\sigma(t_{XY})$ or $\text{APD}(t_{XY})$ would capture participants' judgments better than the Delay_I posterior. Furthermore, we predict that most participants would base their judgments on the variance of t_{AB} rather than t_{SB}, given that it's easier to estimate the interval between subsequent events, rather than separated events. After the summary we hypothesized that participants' judgments would become more normative, that is, closer to the predictions of Delay_I.

Modeling all participants. To establish which combination of these measures best explains participants' judgments we entered them all into a competitive, stepwise, model selection procedure. We used all the data for the model selection. As before we fit mixed effect models with random effects for devices within participants. The independent variables were first z-scored meaning that the final beta weights can be interpreted as percentage increase in assignments to the chain for a $1SD$ increase in the value of each independent variable. We entered the following predictors:

Intercept: Positive value captures overall preference for chain, negative for fork.

$\mathbf{cor(t_{SA}, t_{SB})}$: The correlation between the delays $t^{1:k}_{SA}$ and $t^{1:k}_{SB}$.

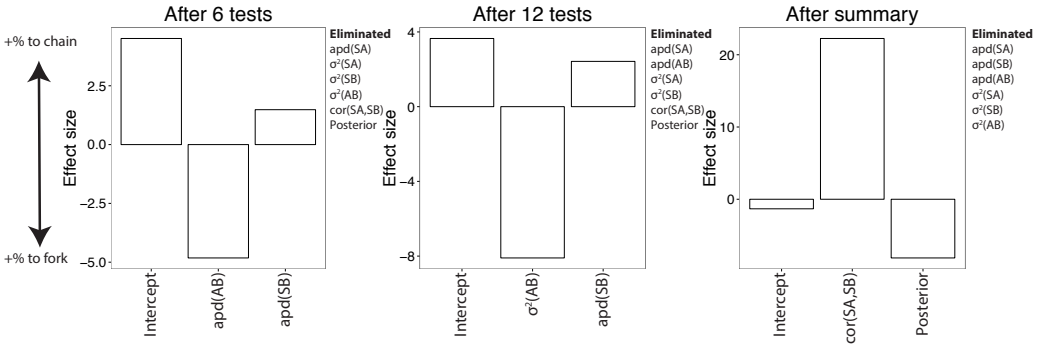

Figure 16. The models resulting from stepwise model selection to all 1040 1st, 2nd and final judgments in Experiment 3. Plots show the selected predictors' fixed effects (i.e., the β values) in descending order of t value. All predictors were z scored, and the dependent variable was centered (so that a prediction of 0 corresponded to assigning 50% to the chain and 50% to the fork). Thus, effect sizes are interpretable as differences in percentage assigned to the chain moving one standard deviation up on the independent variable.

$\boldsymbol{\sigma^2(t_{xy})}$: The variance of the inter-event timing between activation of components x and y in the tests performed so far. We entered the variance for each inter-event interval (i.e., $\sigma^2(t_{SA})$, $\sigma^2(t_{SB})$ and $\sigma^2(t_{AB})$).

$\mathbf{APD}(\boldsymbol{t_{xy}})$: Average pairwise difference. A sequentially computed proxy for variance. The difference in activation time on current test compared to previous test for example t^k_{xy} and t^{k-1}_{xy} summed up over tests $1:K$. E.g. for t_{AB} after six trials this is $\text{APD}(t_{AB}) = \sum_{k=2:6} t^k_{AB} - t^{k-1}_{AB}$. As with the variance, we entered the APD for each inter-event interval.

Posterior: The posterior probability of the chain according to Delay_I

Figure 16 depicts the models selected by the stepwise procedure for the first, second, and third judgments respectively. In all three cases, 2 of the 8 predictors were chosen and the rest eliminated. The chosen predictors were similar for the first two judgments but quite different for the final judgment.

For the first judgment — after 6 tests — participants assigned fewer points to the chain (and more to the fork) if there was high apparent fluctuation in t_{AB}, measured by comparing each test to the previous (i.e., $\text{APD}(t_{AB})$). Fluctuation in t_{SB} was also selected but had a smaller effect in favor of the chain. The fact that the intercept is $\gg 0$ is also suggestive of a baseline preference for the chain that could be overturned by high $\text{APD}(t_{AB})$ or low $\text{APD}(t_{SB})$. The Bayesian posterior was not selected as part of the final model.

For the second judgment — after seeing all the evidence — we see a similar pattern but this time the actual variance of t_{AB} is selected rather than its sequential proxy. Again there is a baseline preference for the chain and a weaker influence of $\text{APD}(t_{SB})$.

For the final judgments — made after seeing the visual summary — we find a different pattern. Now the correlation between t_{SA} and t_{SB} dominates the selected model, so much so that there there is a significant negative relationship with the Delay_I posterior.

We report the correlation between all the predictors in the Appendix.

Summary. In sum, these additional analyses suggest that participants had an initial preference for the chain which was modulated based on their perception of variability in

t_{AB} and, to a lesser extent, in t_{SB}. This is consistent with the idea that many began with an (order-driven) preference for the chain which they could gradually reject if their experienced delays were highly variable under the chain hypothesis. After the visual summary was available, judgments shifted to reflect predominantly the more reliable, but harder to compute, predictive relationship between t_A and t_B — $\text{cor}(t_{SA}, t_{SB})$ — which "popped out" visually when viewing the summary timeline (Figure 14b).

General Discussion

In our first two experiments, we found that people were adept at using order information to make judgments about causal structure, based on a single trial (Experiment 1) and by integrating the information from several observations (Experiment 2). We found that participants generally made the non-simultaneity assumption embodied by our Order_{NS} model, but also distributed likelihood in a way consistent with a preference for similar causal delays within each device. Additional variability could be explained as resulting from uncertainty about how causes combined in the Collider (common effect) structure, some additional sensitivity to the precise event timings, and some degree of conservatism in evidence integration. In Experiment 3, we removed the order cues. In this setting, participants were able to use the variability in the event timings alone to distinguish between a chain and a fork structure. To our knowledge, this is the first time this has been shown experimentally. We now discuss these results more broadly and propose some future directions.

Nonsimultaneity and simultaneity

Like Burns and McCormack (2009) and McCormack et al. (2016), we found the large majority of our participants made judgments in line with a non-simultaneity assumption. This means they considered events that occurred at the same time to be inconsistent with one being caused by the other. However, in parallel they had a preference *for* simultaneity among events that shared a common cause (forks), or effect (colliders), judging it at least as likely that these "common" events will occur simultaneously as one occurring earlier or later. From a continuous time perspective, the probability of perfect simultaneity given variability is strictly zero, while nonsimultaneity, if untied to any particular delays, covers the rest of the space. However, human perception does not have infinite temporal precision. Assuming some perceptual uncertainty, apparent approximate simultaneity is the most likely outcome, with likelihood falling away the greater the perceived discrepancy between the outcomes in either direction.

Causal perception

We looked at only a narrow range of time intervals in the current studies, with trials never lasting more than around three seconds. Weber's law (1834) states that perceptual estimation errors normally grow in proportion to the quantities involved. However, this is known to break down for short (<1 second) intervals which are tracked differently by the brain (Karmarkar & Buonomano, 2007). For short intervals, existing causal beliefs have been shown to shape, or distort, perception (Buehner & Humphreys, 2009; Haggard, Clark, & Kalogeras, 2002), sometimes even leading to reordering of a surprising series of events to

a more "normal" causal order (Bechlivanidis & Lagnado, 2016). This suggests that at this temporal grain, experience is still somewhat under construction (Dennett, 1988), scaffolded by preexisting expectations about causal structure. This also suggests an explanation for why participants in the current experiments sometimes seemed to retain some preference for devices that should have been ruled out (as captured by our ϵ parameter in Experiment 2).

Having formed an impression that a device has a certain structure, someone might easily misperceive a subsequent observation as consistent even if they would usually consider it inconsistent with that structure. This might occur more often if the distortion required to make it consistent is very small. In particular, the simultaneous events that people considered to rule out causation most of the time, might also have been susceptible to being perceived as occurring in the expected causal order. These sorts of effects are not captured by our Order and Delay models which work at Marr's (1982) computational level, and are intentionally scale invariant. However, an interesting project would be to construct a cognitive model that exhibits these patterns. Related to this, a fundamental reason to expect different learning at different timescales comes from the so-called "now or never bottleneck" (Christiansen & Chater, 2016) inherent to experiencing events in real time. When observing closely spaced events, there is little time for explicit comparison of possible structures, or to do anything much beyond constructing an impression of what happened or measuring how wrong your prediction was. Reasoning about relationships between events that are separated by minutes or hours is likely a very different process, as there would be far more time to explicitly reason about and compare hypotheses.

Modality

In the current tasks we looked only at the visual modality. However, it could be that other modalities are even better at inferring patterns in time, audition being an obvious example. Humans (and many other animals), have a finely developed ear for patterns in time and pitch; allowing us to hear and quickly internalize even complex rhythms and melodies (London, 2012). Furthermore, the brain can detect auditory pattern amongst noisy background and even decompose it into its constituents elements (i.e., distinguishing the different instruments in a band). It seems plausible that we evolved these capacities in part to support the search for the reliable patterns in nature that are often clues to its underlying structure (Sloman, 2005). Supporting the notion that the visual modality is better at spotting spatial rather than temporal patterns, we saw that participants were able to make much stronger judgments in Experiment 3 once they saw a visual summary. The summary replaces temporal distance with spatial distance, and suddenly the reliability of t_{AB} pops out clearly (as in Figure14b). During the trials themselves, this realization depended on effortful memorization and comparison across observations.

Conjunctive influence

In Experiments 1 and 2, we instructed participants that the Collider structure was conjunctive — that is, it required both of its causes to activate before the effect would activate. We also included a comprehension question to check that participants had understood this. Nevertheless, around a quarter of participants across both experiments, appeared to treat the Collider as disjunctive (or at minimum not rule out that it was capable of behaving

disjunctively sometimes), assigning nonzero probabilities to the Collider even after observing clips where one of its cause components occurred after the effect, or nonzero likelihoods for the Collider to patterns with only one cause occurring before the effect. This suggests that people default to the disjunctive assumption so strongly that it can either overrule instructions, or fill in if the instructions were forgotten (cf. Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Lucas & Griffiths, 2010; Yeung & Griffiths, 2015).

Additionally, people might have struggled to make sense of the idea of a conjunction in the context of the abstract tasks they were solving. Indeed, formalizing the conjunction for our Delay models forced us to think about what would be a plausible mechanism. Concretely, we assumed that the earlier-arriving causal influence waited around in a buffer for the latter to arrive. However, it would have also been plausible to assume that the two causes have influences that must (at least approximately) coincide in their arrival time in order for a threshold to be reached that triggers the effect. Additionally, people might find conjunctive influence a more natural in situations where at least one of the causal relationships has a sustained or continuous effect (e.g., so that the second event simply tips the level of influence over a threshold that causes the activation of the effect). In general, participants were more uncertain about devices where the impact of the evidence depended on assumptions about how the Collider worked. Given the ambiguity about the exact way in which the Collider worked, participants' increased uncertainty for situations involving these cases may be considered a rational response.

The blessing of variability

Our experimental design highlights an interesting and counterintuitive property of temporal causal inference. Unreliable systems can actually be simpler to uncover. The more unreliable the timings of the events are, the more frequently revealing order reversals will occur, and the more a learner can rely on simple qualitative Order inference. A similar principle applies in the absence of revealing order information. It is actually the variability in delays that provides the signal that our Delay models use to infer the generative causal structure. If the causal delays are perfectly reliable it becomes impossible to distinguish between the order-consistent structures based on their timing.[18] This has interesting parallels to the case of learning from contingency information. In a deterministic system, chains and forks are indistinguishable from contingencies because both effects always covary with their root cause. However, they can be covariationally distinguished in various settings provided the relationships are at least a little unreliable (Bramley, Dayan, & Lagnado, 2015; Fernbach & Sloman, 2009).

Toward a process model

Some participants in Experiments 1 and 2 formed preferences for structures that rendered causal delays more similar on average across connections and clips (reflected by the shift across clips 2 to 6 in Experiment 1, and the few individuals better described by our time-sensitive Delay_P than our qualitative Order models). Additionally, participants'

[18] Assuming you do not have a prior expectation about the lengths of the different delays. Of course, structures could still be distinguished without variability if you know how long the links should take to work.

distribution over qualitative patterns was highly consistent with a preference for equal delays. Judgments in Experiment 3 were consistent with the proposal that people tend to "see" the evidence through the lens of one causal model at a time (Bechlivanidis, 2015), becoming more likely to switch if observed events are sufficiently hard to accommodate under this presumptive structure (Bonawitz et al., 2014; Bramley et al., in revision). Since seeing several events that always occur in the same order *ceteris paribus* is most naturally percieved as a chain, participants may have begun the problems in Experiment 3 with a sense of watching a causal chain, which could be gradually overturned in the cases where there was another more predictive perception available (of the device as a fork). More generally, by pulling these ideas together, we get a picture of temporal causal structure learning in which learners have an initial impression of causal structure based on event order (Lagnado & Sloman, 2006) but are capable of refining this as they observe more evidence about the system and consider what structural changes from this default might make the event times more predictable.

Building richer causal representations

While CBNs provide our current best framework for building theories about causal cognition, they are not rich enough to explain central aspects of causal cognition such as mechanism knowledge and mental simulation (Mayrhofer & Waldmann, 2015; Sloman & Lagnado, 2015; Waldmann & Mayrhofer, 2016) or to ground everyday causal judgments (Gerstenberg et al., 2015). People's causal representations almost certainly lie somewhere in between a compact statistical map (a CBN) and a scale model of the physical world. We can often get away with treating detailed mechanisms as black boxes (Keil, 2006), but we still need our representation to help us choose when and where to act in the world. Thus, it seems necessary that people's representations sometimes include expectations about delays between causes and effects. Of course our causal representation of the world is rich in space as well as time, with detailed knowledge of mechanisms likely to be intertwined with delay expectations. Our generative Delay models represent a step toward capturing the ways in which human causal cognition goes beyond statistical contingencies.

Conclusions

In conclusion, we have shown in three experiments that people form clear and sensible beliefs about causal structure based on temporal information. We can capture people's inferences with a combination of qualitative order-based, and generative delay-based inference models. Participants were able to use the order in which events occurred to narrow in on candidate causal structures, and within these, favored those that rendered the causal delays more similar and more predicable. Going beyond order patterns, we showed that people can also use interval variability alone to identify whether a structure is a chain or a fork, and proposed how participants might achieve this while "seeing" the evidence through the lens of one hypothesis at a time. These results contribute to understanding of the role of time in causal learning and representation, showing that just as time is inherent to our experience of the world, it is integral to our causal models of the world.

Acknowledgments

We thank Christos Bechlivanidis, Michael Pacer, Toby Pilditch, and Steven Sloman for helpful comments. NB is supported by a 4-year Engineering and Physical Sciences Research Council UK stipend awarded by The UCL Centre for Doctoral Training in Financial Computing and Analytics. TG is supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216, and by an ONR grant N00014-13-1-0333. RM is supported by a grant from the Deutsche Forschungsgemeinschaft (Ma 6545/1-2) as part of the priority program "New Frameworks of Rationality" (SPP 1516). DL is supported by an Economic and Social Research Council UK grant (RES 062330004).

References

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.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Bechlivanidis, C. (2015). The arrow of time through the causal lens: When causal beliefs determine temporal order. *Unpublished PhD thesis*.

Bechlivanidis, C., & Lagnado, D. A. (2013). Does the "why" tell us the "when"? *Psychological Science*, *24*(8), 1563–1572.

Bechlivanidis, C., & Lagnado, D. A. (2016). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58–66.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive psychology*, *74*, 35–65.

Bramley, N. R., Dayan, P., & Lagnado, D. A. (2015). Staying afloat on Neurath's boat– Heuristics for sequential causal learning. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 262–267). Austin, TX: Cognitive Science Society.

Bramley, N. R., Griffiths, T. L., Dayan, P., & Lagnado, D. A. (in revision). Formalizing Neurath's ship – Approximate algorithms for online causal learning.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Forgetful conservative scholars - How people learn causal structure through interventions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *41*(3), 708-731.

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Courier Corporation.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*, 1119–40.

Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science*, *20*(10), 1221–1228.

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

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology*, *57*(2), 179–191.

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

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3–5.

Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning*, *15*(2), 167–196.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental

constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

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

Corner, A., Harris, A. J., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In *Proceedings of the 32nd annual meeting of the cognitive science society* (Vol. 1625, p. 1630).

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS ONE*, *8*(3), e57410.

Dennett, D. C. (1988). The intentional stance in theory and practice. In R. Byrne & A. Whiten (Eds.), *Machiavellian intelligence* (pp. 180–202). Oxford, UK: Oxford University Press.

Deverett, B., & Kemp, C. (2012). Learning deterministic causal networks from observational data..

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*, 17–52.

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

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a bayesian perspective. *Psychological review*, *90*(3), 239.

Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are causal structure and intervention judgments inextricably linked? A developmental study. *Cognitive Science*, *36*, 261–285.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis. texts in statistical science series.* Chapman & Hall/CRC, Boca Raton, FL,.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Tenenbaum, J. B. (in press). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning.* Oxford University Press.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110-9.

Gopnik, A., & Sobel, D. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental ...*. Retrieved from `http://psycnet.apa.org/journals/dev/37/5/620/`

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. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory and cognition*, *35*(3), 444–453.

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

Greville, W. J., Cassar, A. A., Johansen, M. K., & Buehner, M. J. (2013). Structural awareness mitigates the effect of delay in human causal learning. *Memory and cognition*, 1–13.

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. (2005). Causes, coincidences, and theories. *Unpublished doctoral dissertation*.

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. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature neuroscience*, *5*(4), 382–385.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory and Cognition*, *30*(7), 1128–1137.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(06), 803–831.

Hartigan, J. A. (2012). *Bayes theory*. Springer Science & Business Media.

Hauser, D. J., & Schwarz, N. (2015). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 1–8.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

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

Hume, D. (1748/1975). *An enquiry concerning human understanding*. Oxford University Press.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., ... Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron*.

Karmarkar, U. R., & Buonomano, D. V. (2007). Timing in the absence of clocks: encoding time in neural network states. *Neuron*, *53*(3), 427–438.

Keil, F. C. (2006). Explanation and understanding. *Annual review of psychology*, *57*, 227.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive science*, *34*(7), 1185–243.

Kuhn, G., Caffaratti, H. A., Teszka, R., & Rensink, R. A. (2014). A psychologically-based taxonomy of misdirection. *Frontiers in psychology*, *5*, 1392.

Lagnado, D. A. (2011). Thinking about evidence. In *Proceedings of the british academy* (Vol. 171, pp. 183–223).

Lagnado, D. A., & Gerstenberg, T. (in press). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning*. Oxford University Press.

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. In W. Gray & C. D. Schunn

(Eds.), *Proceedings of the 24th annual meeting of the cognitive science society.* NJ: Erlbaum.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(3), 451–60.

Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*(2), 184–195.

Lagnado, D. A., Waldmann, M., Hagmayer, Y., & Sloman, S. (2007). Beyond covariation: cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (p. 154-72). London: Oxford University Press.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015, dec). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in cognitive sciences*, *9*(2), 75–82.

London, J. (2012). *Hearing in time: Psychological aspects of musical meter.* Oxford University Press.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955.

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

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Marr, D. (1982). *Vision.* New York: Freeman & Co.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, *44*(1), 1–23.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive science*, *39*(1), 65–95.

Mayrhofer, R., & Waldmann, M. R. (in press). Sufficiency and necessity assumptions in causal structure induction. *Cognitive science.*

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1-22.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277.

Michotte, A. (1946/1963). *The perception of causality.* Basic Books.

Pacer, M. D., & 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* (Vol. 1, pp. 833–838).

Pacer, M. D., & Griffiths, T. L. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In *Proceedings of the 37th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press (2nd edition).

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review*, *14*(4), 577–596.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: a memory-based anchor model of category rating and absolute identification. *Psychological Review*, *112*(2), 383.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior research methods*, *47*(2), 309–327.

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*.

Rottman, B. M., & Hastie, R. (2016, jun). Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134.

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

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*, 303–317.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. *The psychology of learning and motivation*.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* Oxford: Oxford University Press.

Sloman, S. A., & Lagnado, D. (2005). Do we "do"? *Cognitive Science*, *29*(1), 5-39.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247.

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

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Waldmann, M., & Mayrhofer, R. (2016). Hybrid causal representations. *Psychology of Learning and Motivation*.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae.* C F Koehler.

White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, *18*(3), 454–480.

Wolfe, J. B. (1921). The effect of delayed reward upon learning in the white rat. *Journal of Comparative Psychology*, *17*(1), 1.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. In *Psychology of learning and motivation* (pp. 167–202). Elsevier.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

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

Appendix

Collider likelihood

Pooled model. For the Collider, event E happens as the two causal influences of A and B arrived (i.e., conjunctive common-effect; see Equation 4). Thus, the observed between-event intervals t_{AE} and t_{BE} may contain waiting time and so do not necessarily reflect the underlying causal delays $t_{A\to E}$ and $t_{B\to E}$ as we have assumed for the other structures. To model the joint likelihood of the two observed intervals, we have to discriminate two cases: Either (1) the causal influence of B was waiting for the influence of A and therefore E happened as the delay of A arrived (i.e., $t_{AE} = t_{A\to E}$ but $t_{BE} \geq t_{B\to E}$) or (2) the causal influence of A was waiting for the influence of B to arrive and E happened as the delay of B arrived (i.e., $t_{BE} = t_{B\to E}$ but $t_{AE} \geq t_{A\to E}$).

Let the influence of B waiting for A (i.e., Case 1). In this case, the joint likelihood is given by the gamma likelihood of t_{AE} (as t_{AE} does in fact equal $t_{A\to E}$ and is therefore gamma distributed) weighted by the probability of t_{BE} being in fact larger than the respective gamma distributed event $t_{B\to E}$. As we assume the same parameters α and μ for both links (pooled model), the likelihood can be written as

$$p(t_{AE}, t_{BE}|\alpha, \mu; t_{AE} = t_{A\to E}, t_{BE} \geq t_{B\to E}) = p(t_{AE}|\alpha, \mu) \cdot p(t_{BE} \geq t_{B\to E}|\alpha, \mu) \quad \text{(A-1)}$$

Analogously, for the case in which A is waiting for B (i.e., Case 2) it holds

$$p(t_{AE}, t_{BE}|\alpha, \mu; t_{AE} \geq t_{A\to E}, t_{BE} = t_{B\to E}) = p(t_{BE}|\alpha, \mu) \cdot p(t_{AE} \geq t_{A\to E}|\alpha, \mu) \quad \text{(A-2)}$$

As both cases are mutual exclusive and therefore constitute a partitioning of the joint likelihood, the joint likelihood can be written as a sum of both (law of total probability)

$$p(t_{AE}, t_{BE}|\alpha, \mu) = p(t_{AE}, t_{BE}|\alpha, \mu; t_{AE} = t_{A\to E}, t_{BE} \geq t_{B\to E}) + p(t_{AE}, t_{BE}|\alpha, \mu; t_{AE} \geq t_{A\to E}, t_{BE} = t_{B\to E}) \quad \text{(A-3)}$$

$$= p(t_{AE}|\alpha, \mu) \cdot p(t_{BE} \geq t_{B\to E}|\alpha, \mu) + p(t_{BE}|\alpha, \mu) \cdot p(t_{AE} \geq t_{A\to E}|\alpha, \mu) \quad \text{(A-4)}$$

with $p(t_{AE}|\alpha, \mu)$ and $p(t_{BE}|\alpha, \mu)$ being gamma distributed and $p(t_{AE} \geq t_{A\to E}|\alpha, \mu)$ and $p(t_{BE} \geq t_{B\to E}|\alpha, \mu)$ following the gamma's cumulative distribution function with

$$p(t_{AE} \geq t_{A\to E}|\alpha, \mu) = \int_0^{t_{AE}} \frac{(\frac{\alpha}{\mu})^\alpha}{\Gamma(\alpha)}(x)^{\alpha-1} e^{-\frac{\alpha}{\mu}x}\, dx \quad \text{(A-5)}$$

and for $p(t_{BE} \geq t_{B\to E}|\alpha, \mu)$ analogously.

Independent model. In the independent model, each causal connection between a variable X and its effect Y is assumed to have its own set of parameters α_{XY} and μ_{XY}. Therefore, the Collider likelihood in the independent model is given by

$$\begin{aligned} p(t_{AE}, t_{BE}|\alpha_{AE}, \alpha_{BE}, \mu_{AE}, \mu_{BE}) =& p(t_{AE}|\alpha_{AE}, \mu_{AE}) \cdot p(t_{BE} \geq t_{B\to E}|\alpha_{BE}, \mu_{BE}) \\ &+ p(t_{BE}|\alpha_{BE}, \mu_{BE}) \cdot p(t_{AE} \geq t_{A\to E}|\alpha_{AE}, \mu_{AE}) \end{aligned} \quad \text{(A-6)}$$

Disjunctive Collider. In our experiments, we used conjunctive Colliders. However, in other scenarios a disjunctive combination function of the causal influences may be more natural. In this case, the activation time of effect event E is determined by the first arrival of the causes' influences

$$t_E = \min[t_A + t_{A\to E}, t_B + t_{B\to E}] \tag{A-7}$$

In this case, one of the underlying causal delays $t_{A\to E}$ or $t_{B\to E}$ is overshadowed by E's happening resulting in a smaller observed delay. Analogously to the conjunctive Collider, there are two cases: (1) the influence of A arrives first, causing E to happen and overshadowing the influence of B (i.e., $t_{AE} = t_{A\to E}$ but $t_{BE} \leq t_{B\to E}$) and (2) the influence of B arrives first overshadowing the influence of A (i.e., $t_{BE} = t_{B\to E}$ but $t_{AE} \leq t_{A\to E}$). Thus, the joint likelihood of a disjunctive (pooled delay) Collider can be written as

$$p(t_{AE}, t_{BE}|\alpha,\mu) = p(t_{AE}|\alpha,\mu)\cdot p(t_{BE} \leq t_{B\to E}|\alpha,\mu) + p(t_{BE}|\alpha,\mu)\cdot p(t_{AE} \leq t_{A\to E}|\alpha,\mu) \tag{A-8}$$

$$= p(t_{AE}|\alpha,\mu)\cdot(1 - p(t_{BE} \geq t_{B\to E}|\alpha,\mu)) + p(t_{BE}|\alpha,\mu)\cdot(1 - p(t_{AE} \geq t_{A\to E}|\alpha,\mu)) \tag{A-9}$$

Simple Monte Carlo — Experiments 1 and 2

As there is no closed form solution for the marginal likelihoods $p(\mathbf{d}\,|s)$ of data $\mathbf{d}$ under structure s, we used a simple Monte Carlo sampling scheme to approximate the multiple integral. For this purpose, we drew $B = 100,000$ independent samples from the respective parameters' prior distributions $p(\lambda|s)$, $p(\alpha|s)$ and $p(\mu|s)$ and averaged over the likelihoods (see Equation 9) at the sampled points in parameter space

$$p(\mathbf{d}\,|s) = \int p(\mathbf{d}\,|\lambda,\alpha,\mu;s)\cdot p(\lambda|s)\cdot p(\alpha|s)\cdot p(\mu|s)\; d\lambda\; d\alpha\; d\mu \tag{A-10}$$

$$= \frac{1}{B}\sum_{b=1}^{B} p(\mathbf{d}\,|\lambda^{(b)},\alpha^{(b)},\mu^{(b)};s) \tag{A-11}$$

with $\lambda^{(b)}$, $\alpha^{(b)}$, and $\mu^{(b)}$ being the b's sampled points from the prior distributions.

Markov Chain Monte Carlo estimation — Experiment 3

In Experiment 3, we could use an uninformative prior for the parameters of the gamma distribution (as no collider was involved). For one causal link and the gamma's (α,θ) parametrization with $\mu = \frac{\alpha}{\theta}$, we can derive the posterior based on a conjugate prior assuming "no prior observations"

$$p(\alpha,\theta|\,\mathbf{d};s) \propto \frac{p^{\alpha-1}e^{-\frac{q}{\theta}}}{\Gamma(\alpha)^n\theta^{\alpha n}} \tag{A-12}$$

for n data points $\mathbf{d}$ with $p = \prod d_i$ and $q = \sum d_i$.[19] The normalizing constant of the equation's right hand side is our target of interest, namely the marginal likelihood of the data given the structure of interest $p(\mathbf{d}\,|s)$. To approximate the integral, we used a two-step procedure:

1. We generated a sample from the posterior over α and θ via the Metropolis–Hastings algorithm (i.e., MCMC) with 10,000 points sampled from 10 chains each with Gaussian proposal distribution on α ($SD = 10$) and θ ($SD = 5$) and burn-in of 1,000 and only each tenth point taken (i.e., thinning). We run the sampler ten times to check for convergence (see Gelman, Carlin, Stern, & Rubin, 2004).

2. We used the obtained sample to estimate the marginal likelihood with the method proposed by Chib and Jeliazkov (2001). Although the method formally works with just one sampled point, we used a subset to generate a more stable estimate.We randomly drew 1,000 points from the MCMC sample and took the 50 points with the largest likelihoods in this subsample. For each of these points, we calculated the marginal likelihood estimate with the method proposed by Chib and Jeliazkov (2001) and averaged over these to get our estimate of $p(\mathbf{d}\,|s)$.

Checking sensitivity to priors

We can assess the sensitivity of the Simple Monte Carlo model predictions to prior choices by comparing them to the predictions of the Markov Chain Monte Carlo procedure we used to estimate posteriors in Experiment 3. The Markov Chain procedure gives posterior predictions based on an uninformative "improper" (Hartigan, 2012) prior but cannot be used for the Collider structure in Experiments 1 and 2. We see in Figures 17 and 18 that there is a little sensitivity to choice of priors on α and μ. Particularly, too high a rate for μ leads to an initial preference for shorter delays and hence the chain under which the delays are necessarily shorter. Additionally, too low a rate for either α or μ led to less stable predictions as few samples fall in the range of the true generative model. However, our chosen values of 0.1 for α and 0.0001 for μ make these effects negligible for the range of event timings we consider the current experiments.

[19] Note that we describe delays in terms of their shape α and mean μ in the main text to aid exposition. However, in statistical applications including approximating inference it is more common and more convenient to work with shape and rate θ.

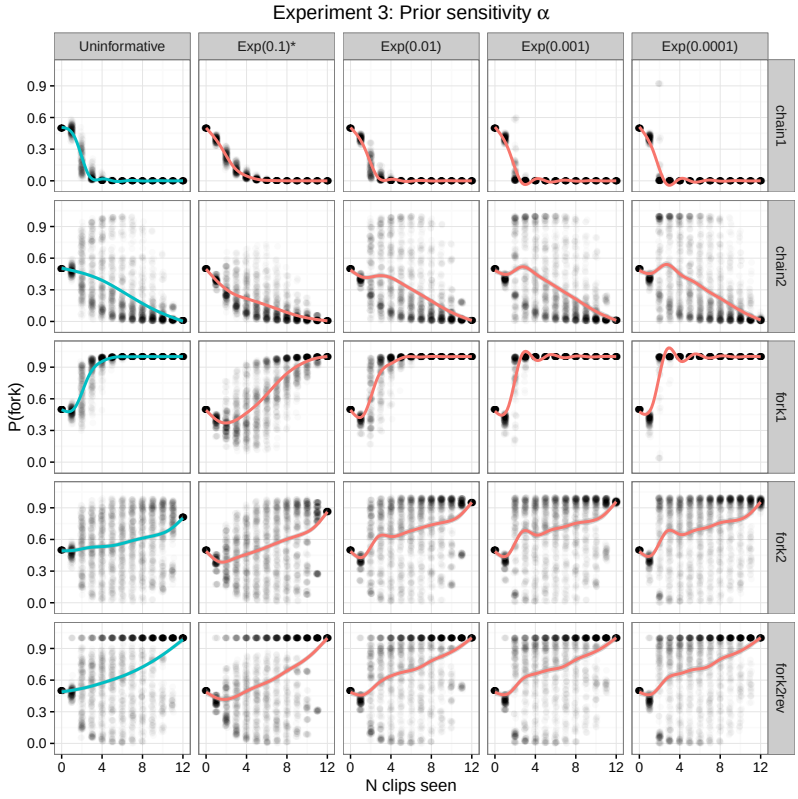

Figure 17. Sensitivity of α prior on model predictions in Experiment 3. Left hand column (teal line) shows predictions using an "improper" uninformative prior. Other columns show predictions under different priors on α. Asterisk indicates the values used for Experiments 1 and 2. The prior on μ for these simulations was Exponential(0.0001). As in Figure 13, individual points are for subsets of the tests seen by different participants at different points during the experiment.

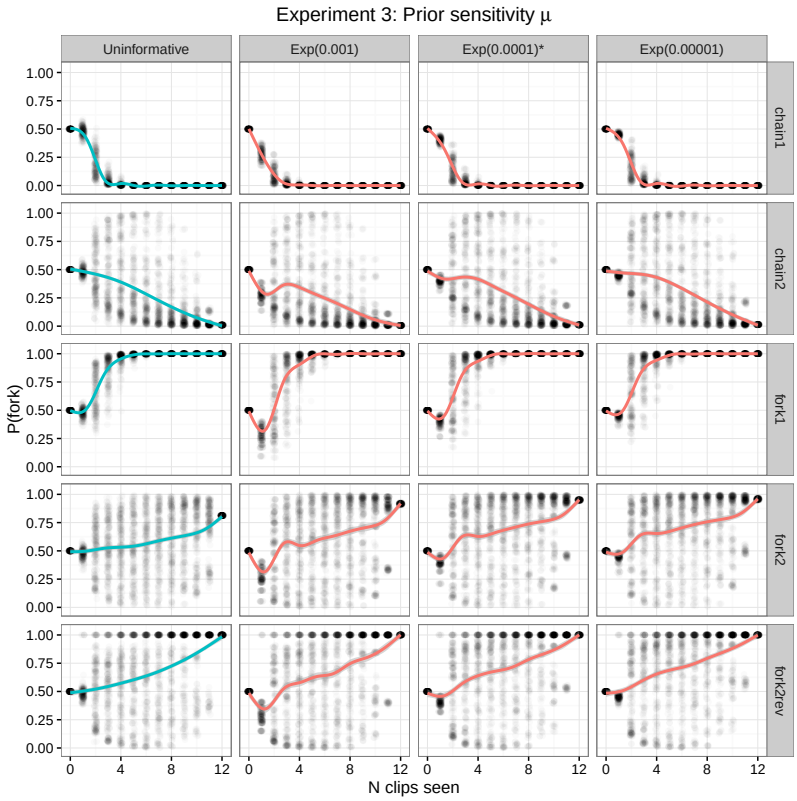

Figure 18. Sensitivity of μ prior on model predictions in Experiment 3. Left hand column (teal line) shows predictions using an "improper" uninformative prior. Other columns show predictions under different priors on μ. Asterisk indicates the values used for Experiments 1 and 2. The prior on α for these simulations was Exponential(0.1).

Correlations Among Heuristic Predictors in Experiment 3

Table 6
Experiment 3: Correlation Between Heuristic Measures

First Judgment (after 6 tests)							
	$\text{cor}(t_{SA}, t_{SB})\sigma(t_{SA})$		$\sigma(t_{SB})$	$\sigma(t_{AB})$	$\text{APD}(t_{SA})$	$\text{APD}(t_{SB})$	$\text{APD}(t_{AB})$
Delay_I	0.69	-0.03	0.42	-0.31	-0.01	0.52	-0.44
$\text{cor}(t_{SA}, t_{SB})$		-0.02	0.13	-0.66	0.01	0.23	-0.68
$\sigma(t_{SA})$			0.15	0.16	0.69	0.12	0.13
$\sigma(t_{SB})$				0.51	0.16	0.90	0.42
$\sigma(t_{AB})$					0.12	0.39	0.88
$\text{APD}(t_{SA})$						0.20	0.17
$\text{APD}(t_{SB})$							0.33
Second and Third Judgments (after all 12 tests)							
	$\text{cor}(t_{SA}, t_{SB})\sigma(t_{SA})$		$\sigma(t_{SB})$	$\sigma(t_{AB})$	$\text{APD}(t_{SA})$	$\text{APD}(t_{SB})$	$\text{APD}(t_{AB})$
Delay_I	0.97	–	0.38	-0.61	0.04	0.45	-0.63
$\text{cor}(t_{SA}, t_{SB})$			0.24	-0.70	0.04	0.37	-0.77
$\sigma(t_{SA})$			–	–	–	–	–
$\sigma(t_{SB})$				0.51	0.03	0.93	0.38
$\sigma(t_{AB})$					-0.01	0.37	0.92
$\text{APD}(t_{SA})$						0.12	0.05
$\text{APD}(t_{SB})$							0.24

Note: In Experiment 3, $\sigma(t_{SA})$ was the same for all devices so it is uncorrelated with all measures once all evidence has been observed.

List of Tables

Table 1
Experiment 1: Order and Delay Models Compared to Participants' Judgments

Model	r	r_s	Mode match	RMSE	N con (N dis)
Baseline				20.1	0
Order_{NS}	0.90	0.76	78%	11.1	12 (10)
Order_{S}	0.71	0.75	56%	16.4	1 (1)
Delay_{P}	0.80	0.64	44%	15.8	3 (4)

Note: Model fits assuming the Collider was conjunctive. r = average Pearson's r correlation between average assignments to structures within each device and model predictions. r_s = average Spearman's rank correlation within problems. Mode match = proportion of problems where participants' modal choice matched model's. RMSE = root mean squared error. N = Number of individuals best correlated by model (con= assuming conjunctive Collider, dis= assuming disjunctive Collider).

Table 2
Experiment 2: Possible Temporal Order Patterns

Pattern	1	2	3	4	5	6	7
Order	$AB \succ E$	$A \succ B \succ E$	$A \succ BE$	$A \succ E \succ B$	$B \succ A \succ E$	$B \succ AE$	$B \succ E \succ A$

Table 3
Experiment 2: Evidence Sets (1^{st} - 4^{th} Piece of Evidence) for the 8 Different Devices

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1st	1	2	2	1	2	2	2	2
2nd	2	2	3	2	2	2	2	2
3rd	5	2	4	5	2	2	2	2
4th	2	2	3	4	4	5	1	6
Shift	N	N	N	Y	Y	Y	Y	Y
Different				N	N	N	Y	Y

Note: The numbers in the rows from 1^{st} to 4^{th} refer to the temporal order patterns shown in Table 2. The roles of components A and B were counterbalanced (e.g., pattern 2 $A \succ B \succ E$ becomes pattern 5 $B \succ A \succ E$) and responses re-coded. *Shift* shows whether a change of MAP judgment is predicted by one or both Order models (N)o/(Y)es. *Different* shows whether this shift is predicted to be different between Order_{NS} and Order_{S}.

Table 4
Experiment 2: Likelihood Judgment Model Fits

Model	r	r_s	Mode match	RMSE	N
Baseline			11%	15.4	1
Order_{NS}	0.92	0.78	71%	8.9	11
Order_{S}	0.57	0.80	43%	12.8	4
Delay_{P}	0.98	0.81	100%	7.3	24

Note: r = average Pearson's r correlation between average assignments to structures within each device and model predictions. r_s = average Spearman's rank correlation within problems. Mode match = proportion of problems where participants' modal choice matched model's. RMSE = root mean squared error. N = Number of individuals best correlated by model.

Table 5
Experiment 3: Main Effects and Planned Comparisons for First, Second and Third Responses

	Response 1	Response 2	Response 3
Main effect (LR)	86***	334***	378***
Planned contrasts			
Intercept	$52 \pm .9\%^{***}$	$46 \pm .9\%^{***}$	$42 \pm .9\%^{***}$
1. Chains vs. Forks	$10.2 \pm 1.3\%^{***}$	$22.2 \pm 1.3\%^{***}$	$33.8 \pm 1.6\%^{***}$
2. Chain1 vs Chain2	$2.7 \pm 1.0\%^{**}$	$4.1 \pm 1.1\%^{**}$	$7.8 \pm 1.3\%^{***}$
3. Fork1 vs Fork2	$2.1 \pm 1.0\%^{*}$	$4.0 \pm 1.1\%^{**}$	$-.6 \pm 1.3\%$
4. Forks1&2 vs Fork2rev	$4.5 \pm 1.1\%^{**}$	$19.6 \pm 1.3\%^{***}$	$16 \pm 1.5\%^{***}$

Note: For main effects we report the likelihood ratio for a model with device type as predictor relative to a model with just an intercept. For each planned comparison we report the size of the effect (%) ± standard error, and level of significance: $^{*} = p < .05$, $^{**} = p < .01$, $^{***} = p < .001$.

List of Figures

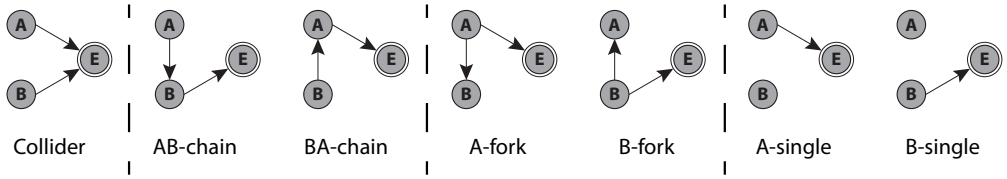

Figure 1. Possible causal structures in Experiments 1 to 3. The arrows indicate the direction of the causal relationship. Dotted lines indicate different types of structure. *Note*: the Collider is conjunctive — both A and B must occur for E to occur.

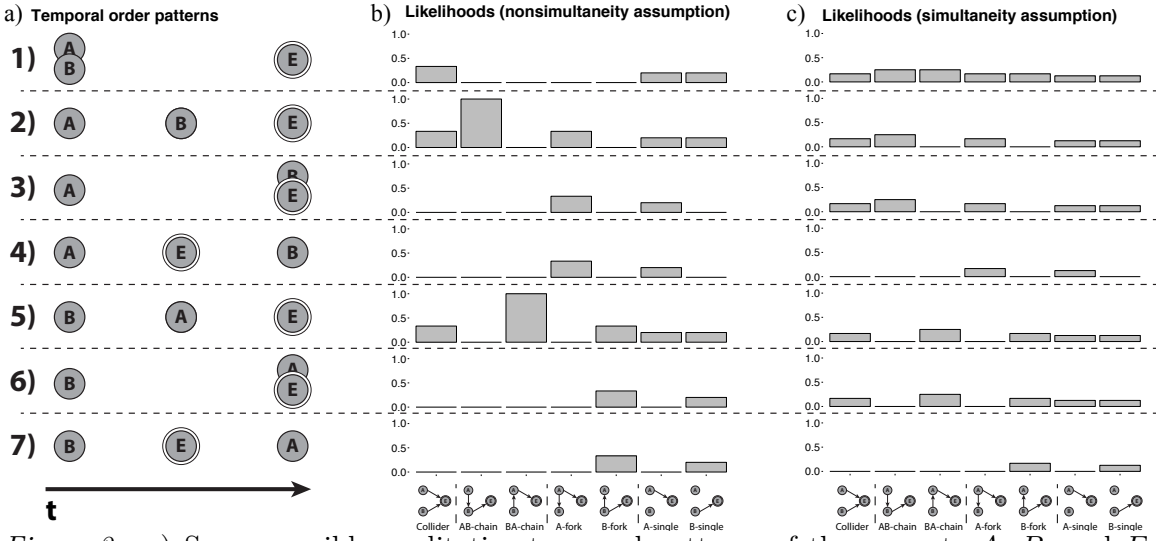

Figure 2. a) Seven possible qualitative temporal patterns of three events A, B, and E. Likelihood functions for the pattern types given the seven different causal structures with non-simultaneity assumption b) or simultaneity assumption c).

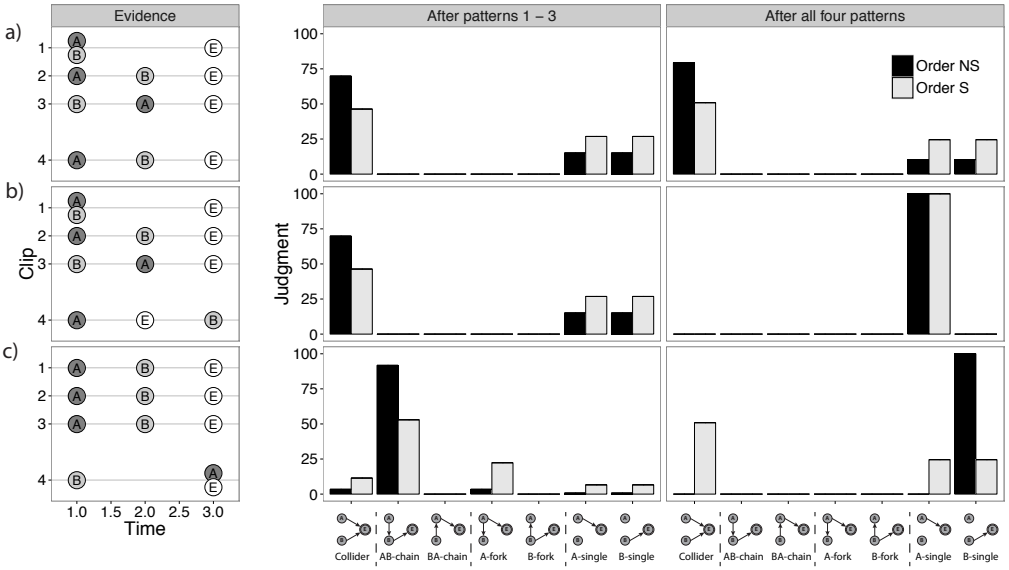

Figure 3. Three examples of order model predictions. Left hand side: Sets of 4 time series showing staggered activation of components A, B and E. Right hand side, model posteriors after seeing clips 1–3 (left column), and after having seen all four patterns (right column).

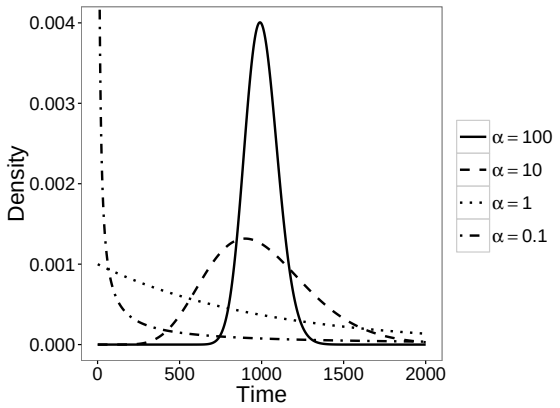

Figure 4. Three example gamma distributions. All have a mean of $\mu = 1000$ ms but differ in their shape α. The exponential distribution is the case where $\alpha = 1$.

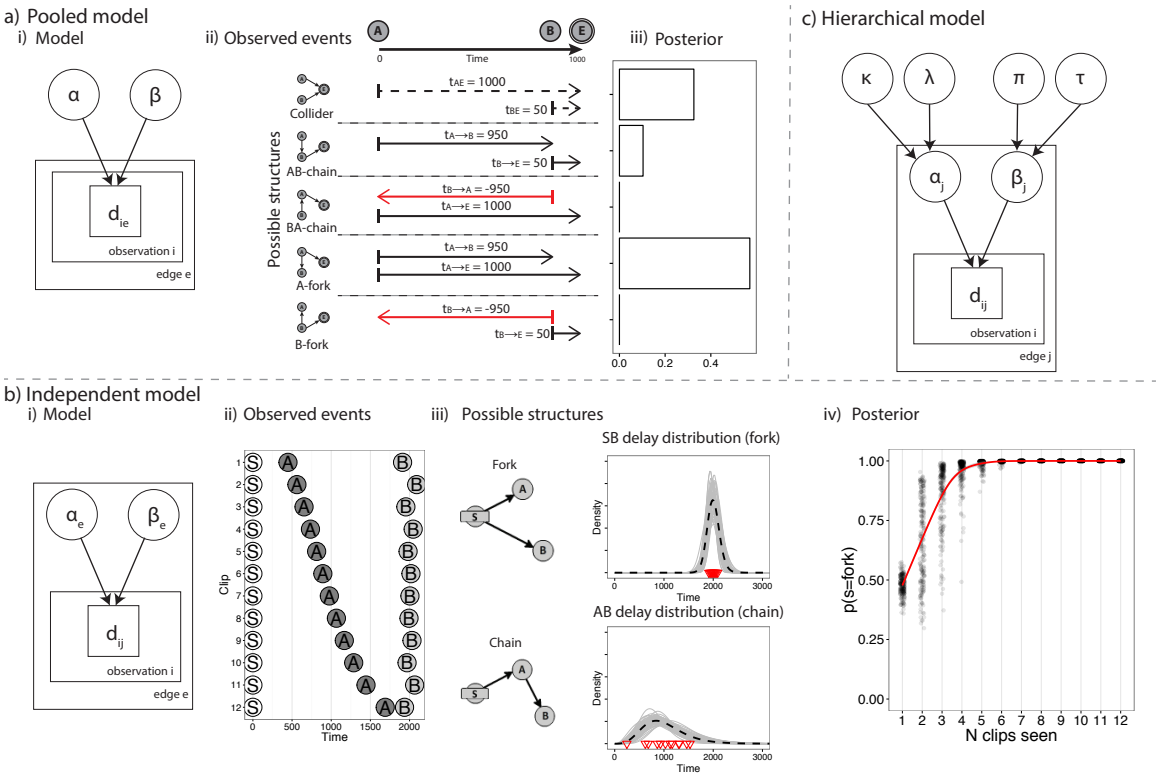

Figure 5. Delay sensitive models and predictions. a) i. *Pooled* Delay_P in plate notation. ii. Example of inference in the pooled model. Observed event timings are mapped onto causal delays under different models. Each row shows the causal delays assuming a different structure. For the Collider, dashed lines indicates that one or other causal delay may be shorter than the observed intervals. Red arrows indicate that structures that can be ruled out based on order alone. iii. Posterior predictions of the delay model assuming priors of $S \sim \text{Unif}(\frac{1}{7})$, $\alpha \sim \text{Exp}(0.1)$, and $\mu \sim \text{Exp}(0.0001)$. b) i. *Independent* Delay_I model in plate notation. ii. 12 patterns of evidence. iii. Posterior marginal inference for two possible structures. The plots show posterior delay samples (gray lines) and their overall density (dotted black line). Both structures share the same $t_{S\to A}$ delays, but the high variance of t_{AB} relative to t_{SB} means this data was more likely produced by a fork as shown in iv., which plots the posterior probability of the fork structure averaged over subsets of the 12 clips (red line gives smoothed average, black dots give posteriors for samples). c) An example of a *hierarchical* Delay_H model in plate notation, where different components have different distributions but are related by hyperparameters.

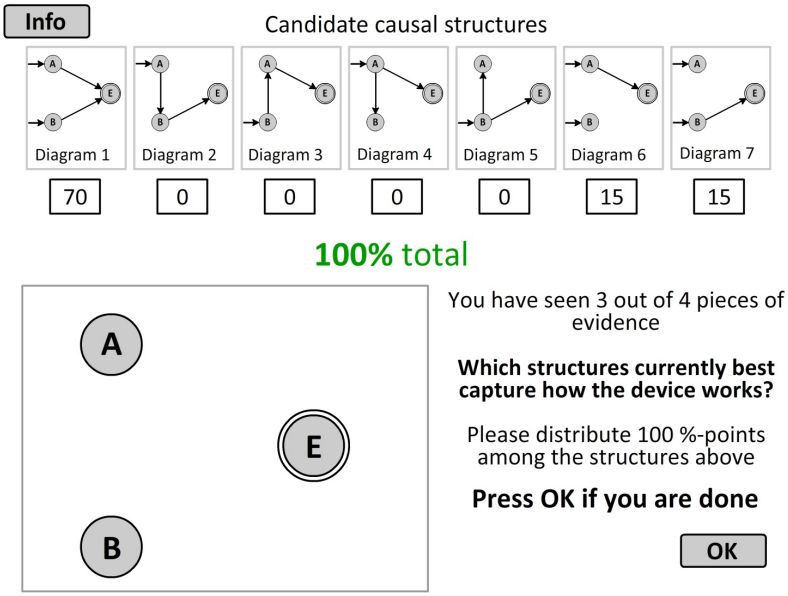

Figure 6. The experiment interface for Experiments 1–3. Clips are shown in the bottom left panel and judgments elicited at the top.

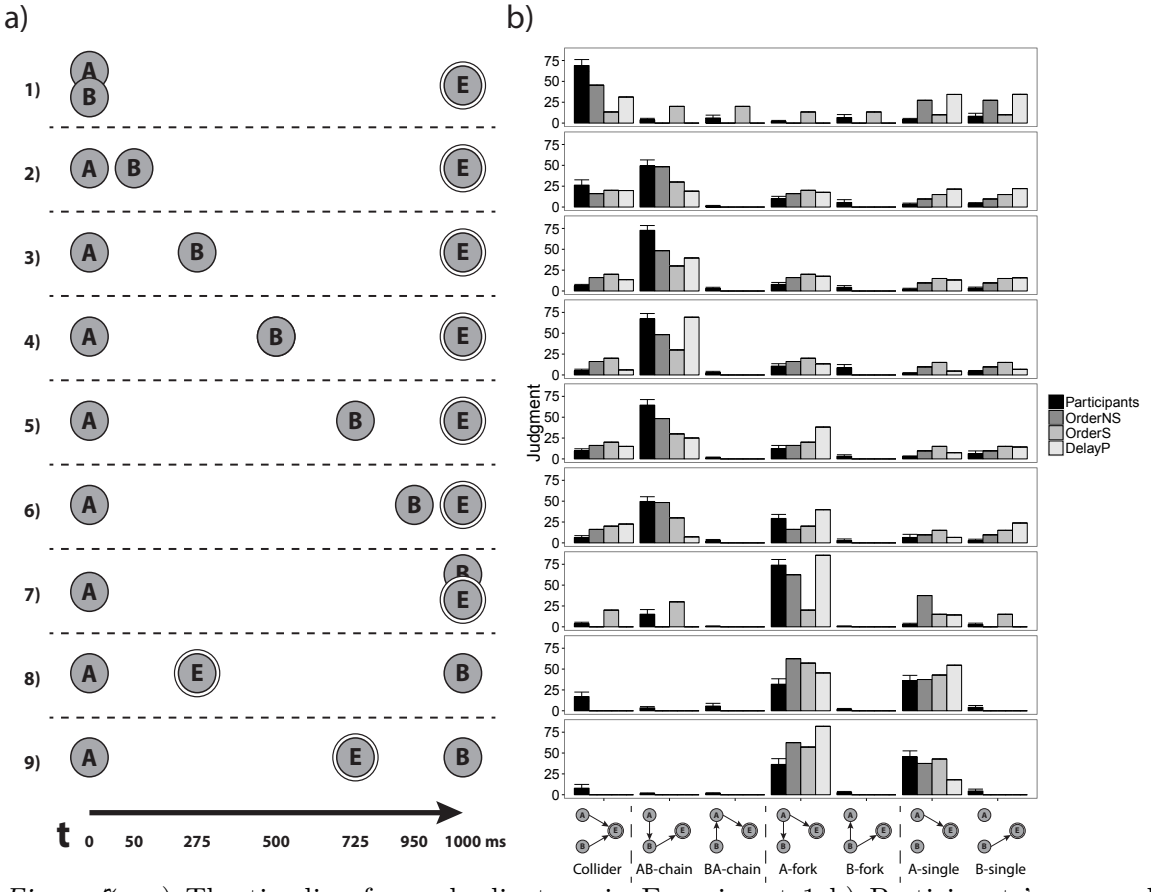

Figure 7. a) The timeline for each clip type in Experiment 1 b) Participants' averaged judgments after viewing each clip (black bars) and predictions by the different models (gray bars). Error bars show standard errors.

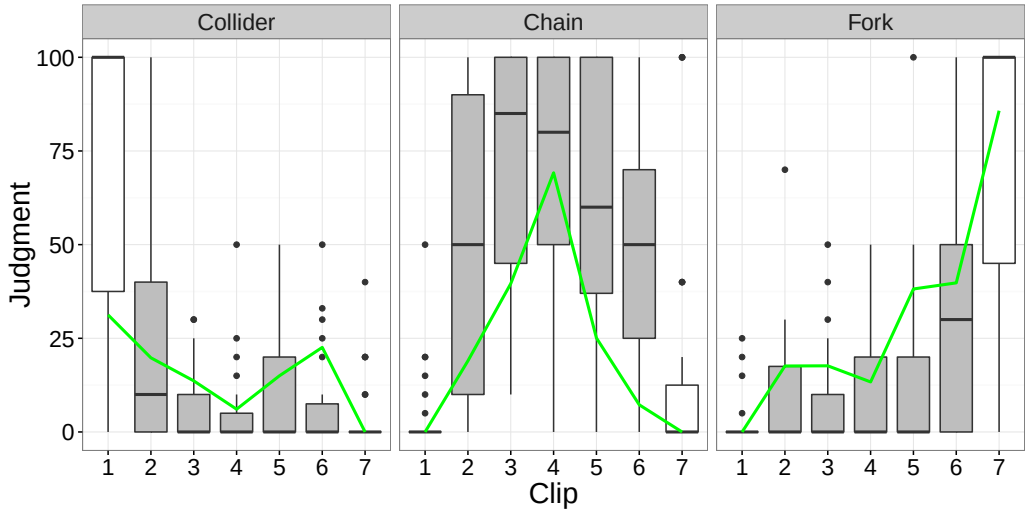

Figure 8. Comparison of probability assignments to Fork, Chain and Collider structures for clips 1–7 (cf. Figure 7), in which B is appears at 0, 50, 275, 500, 725 950 and 1000 ms after A, with E always occurring at 1000 ms. Boxplots show participants' median and upper and lower quartiles, participants with judgments $\pm > 1.5$ interquartile range are plotted separately. Results in text are relative to the six middle bars (gray). Green line denotes Delay_P model predictions.

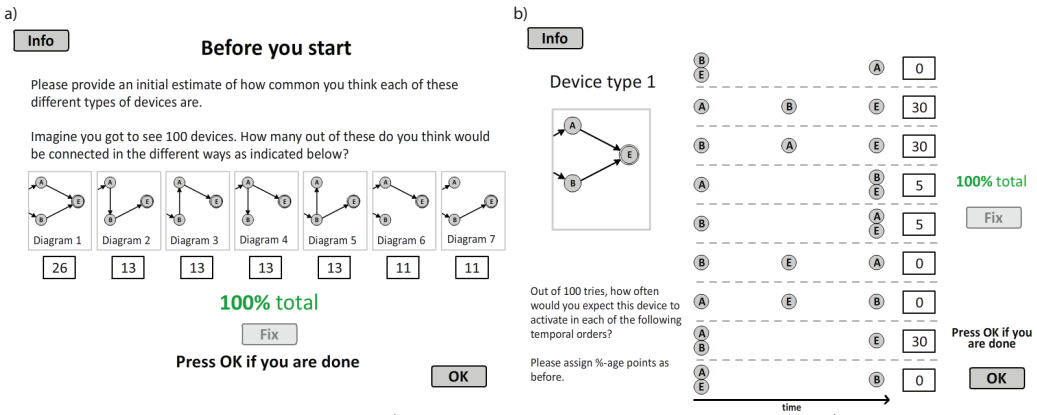

Figure 9. Experiment 2 interface. a) Eliciting priors before main task b) Eliciting likelihoods after main task.

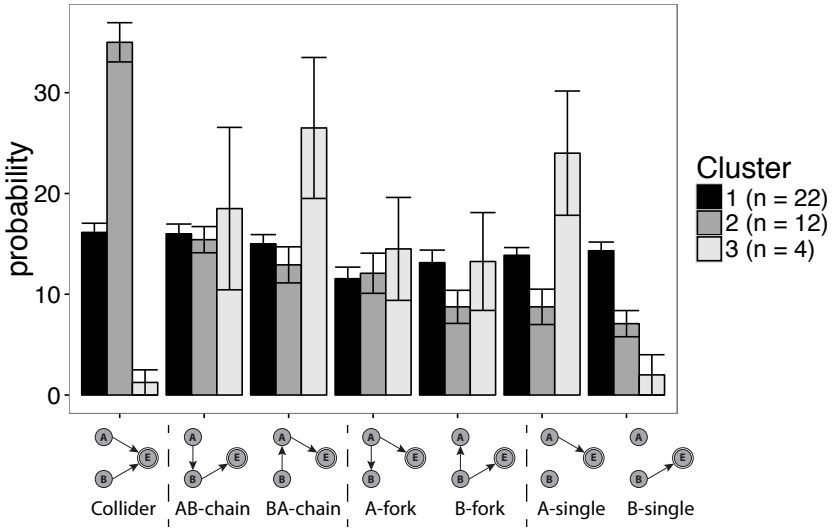

Figure 10. Elicited priors split into clusters as detailed in text. Error bars show standard errors.

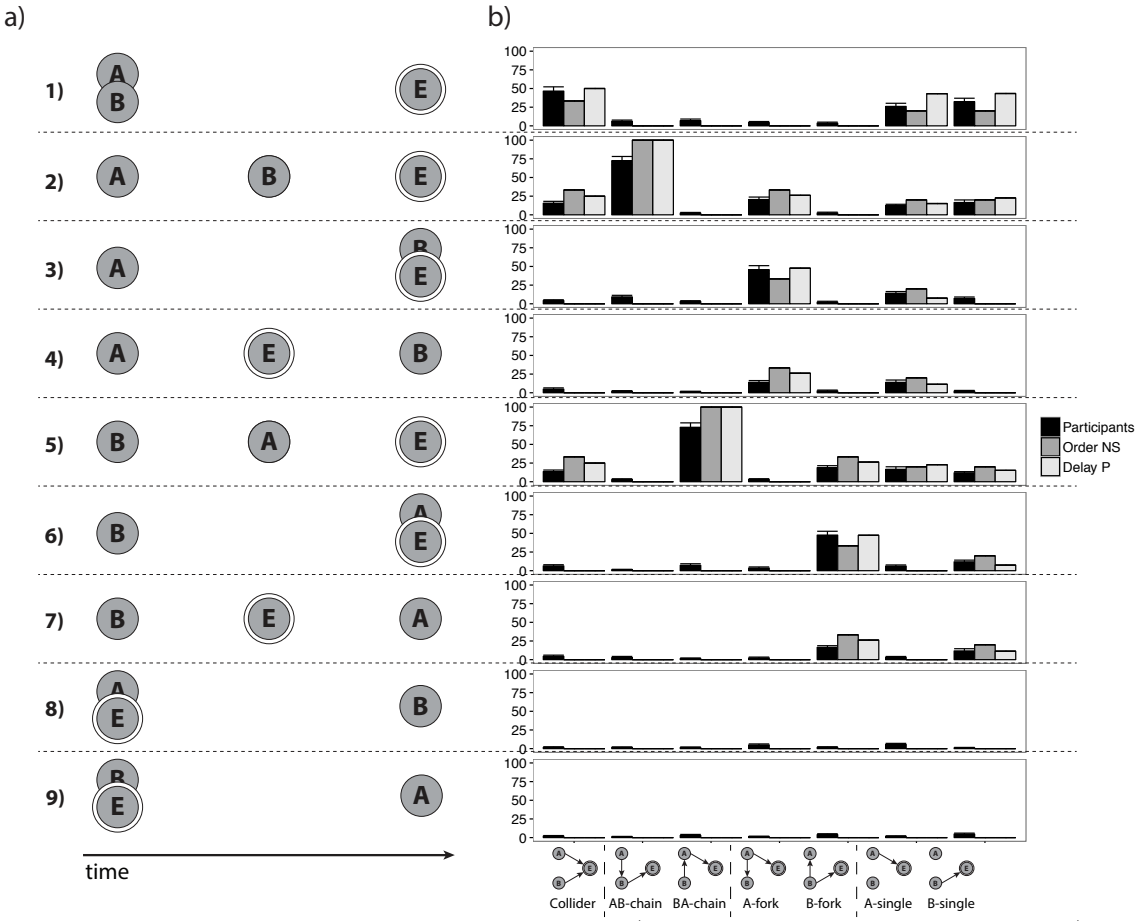

Figure 11. Experiment 2 elicited likelihoods. a) Nine temporal order patterns b) Participants elicited likelihoods compared with those of Order_{NS} and a variant or Order_{NS} type that distributes likelihoods across *types* of consistent patterns — lumping together $A \succ B \succ E$ and $B \succ A \succ E$ for the Collider, and $A \succ B \succ E$ and $A \succ E \succ B$ for the forks. Error bars show standard errors.

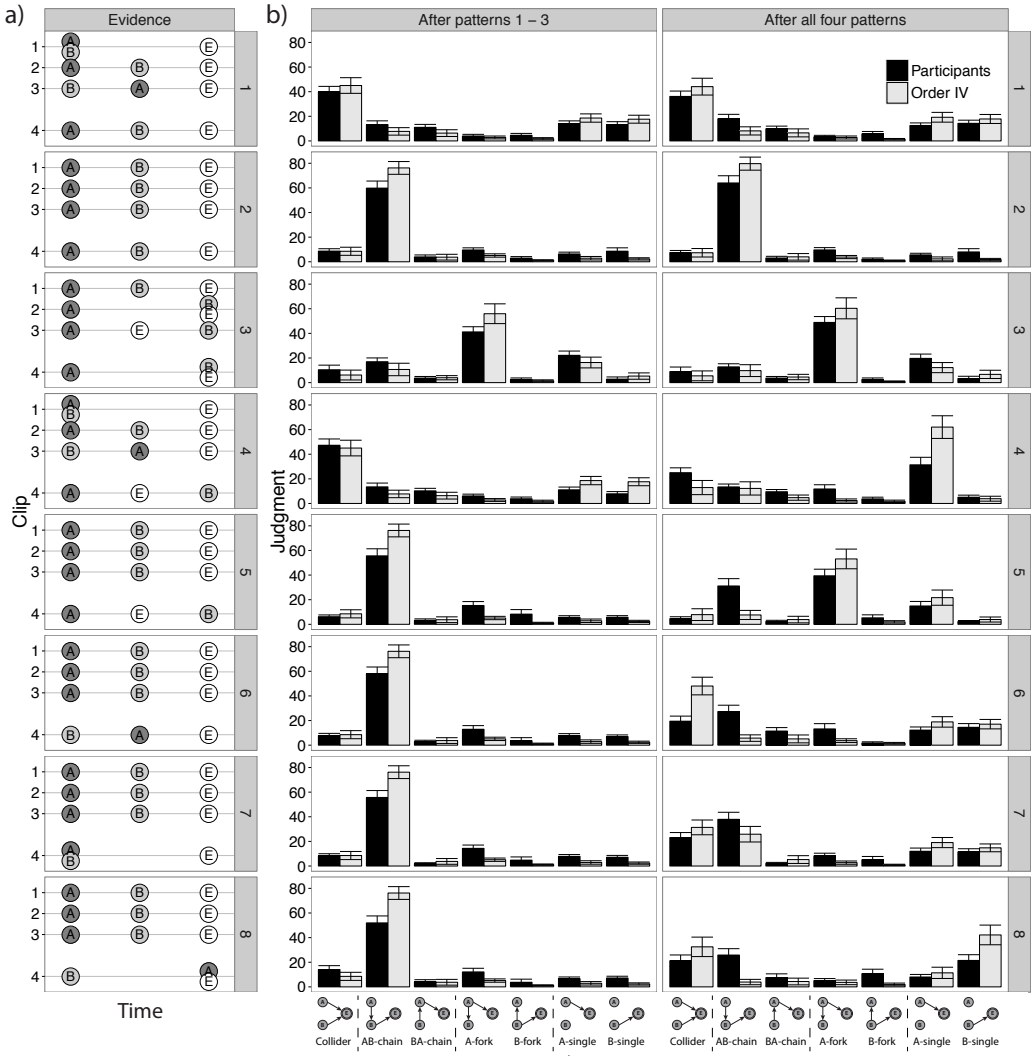

Figure 12. Experiment 2 posterior judgments. a) Devices and qualitative order of activations for each. *Note*: Exact timings were drawn at random from Uniform$(200, 1200)$ for each participant in this experiment and so are not shown in full. b) Participants' posterior judgments (black bars) compared to a model based on individually elicited priors and order-based likelihoods Order_{IV} (gray bars). Left hand column, judgments after viewing 3 clips, right hand column judgments after all four clips. The Order_{IV} bars omit cases in which participants' chosen likelihoods and priors led to all hypotheses being ruled out. Error bars show standard errors.

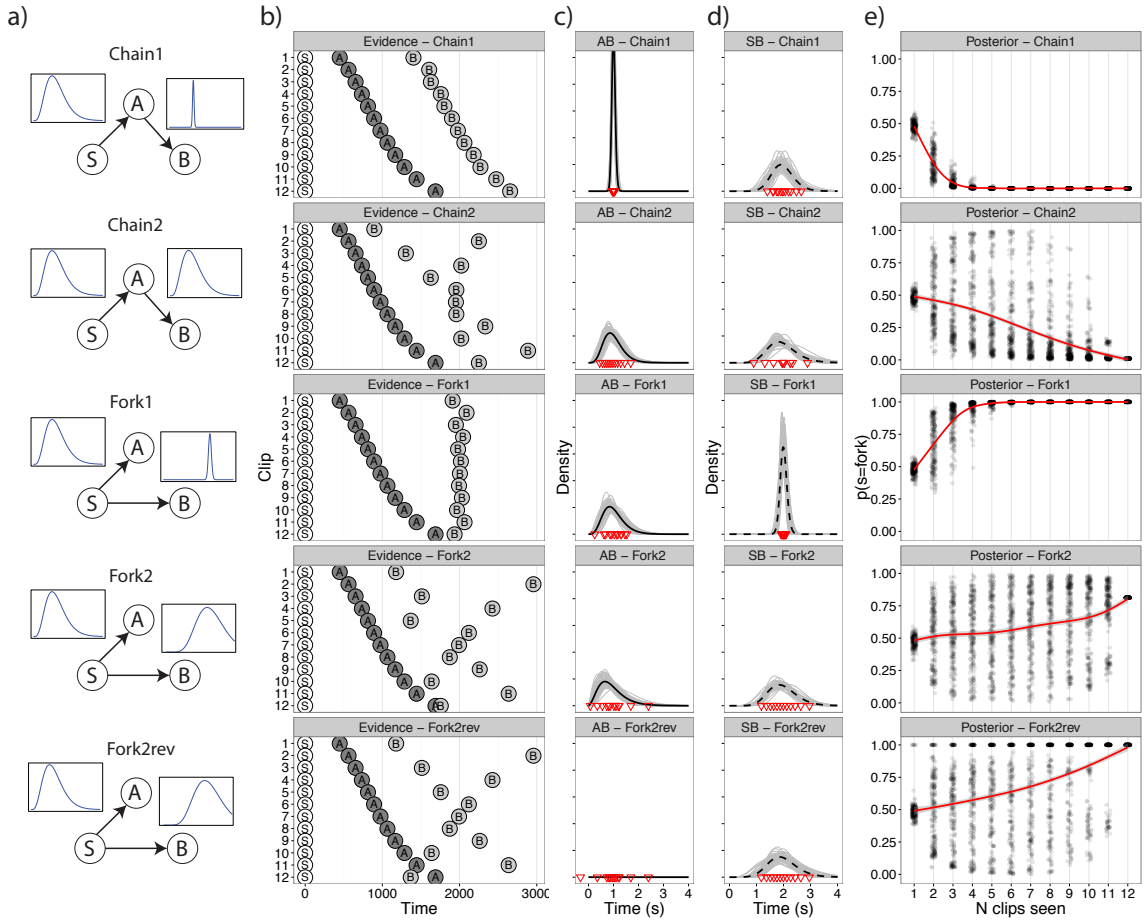

Figure 13. Experiment 3 stimuli and model predictions. a) Graphical representation of the five device types. b) Plot showing the 12 patterns generated for each device. c) Red inverted triangles: t_{AB} for patterns 1:12. Gray lines: $P(G_{A\to B}|\,\mathbf{d})$ for a posterior sample of αs and βs. Dashed black line: The posterior marginal likelihood of $G_{A\to B}$. d) As in c) but for $G_{S\to B}$ under the fork structure. e) Posteriors $P(s = fork|\,\mathbf{d})$ for progressively more evidence. Individual dots for the samples of evidence seen by participants, lines smoothed average (using the general linear additive model with integrated smoothness estimation `gam` from R's `mgcv` library). *Note*: Individual points are jittered along the x-axis to increase visibility.

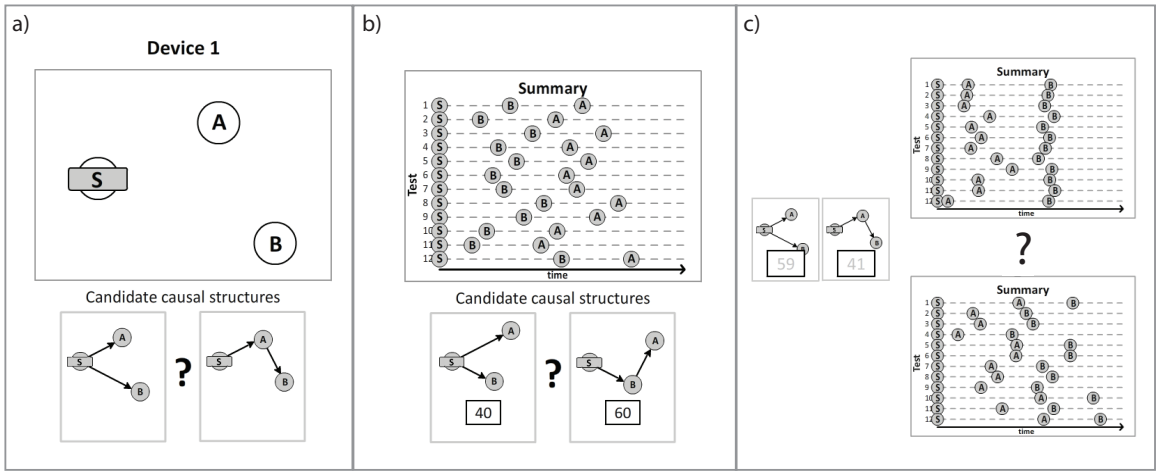

Figure 14. Experiment 3 interface. a) Testing the device b) Viewing a visual summary c) Making a predictive judgment.

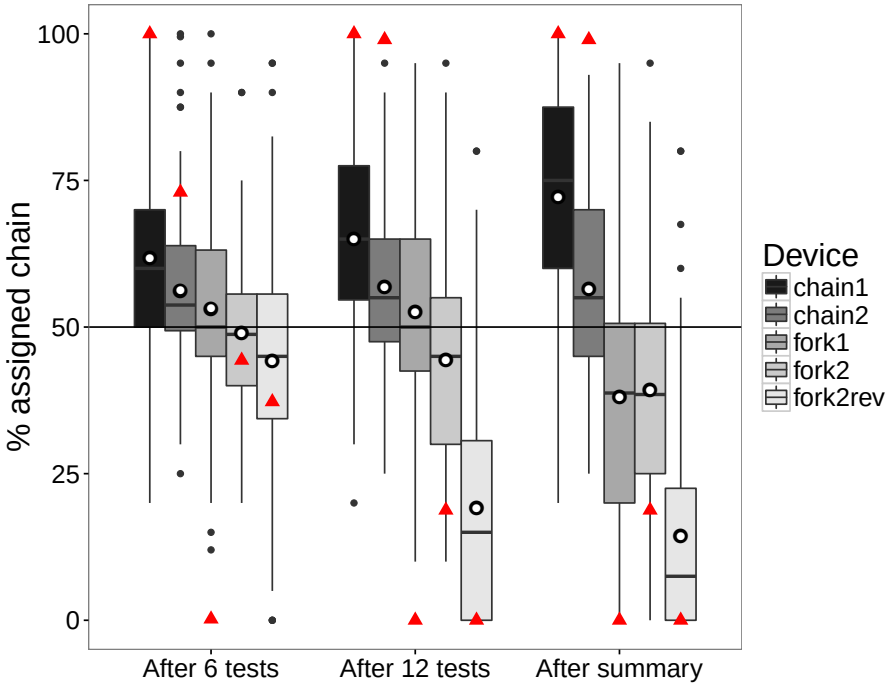

Figure 15. Judgments for the different devices. Boxplots show participants' median and upper and lower quartiles, participants with judgments ± > 1.5 interquartile range are plotted separately. White filled black circles = participant means. Red triangles = Delay_I posteriors.

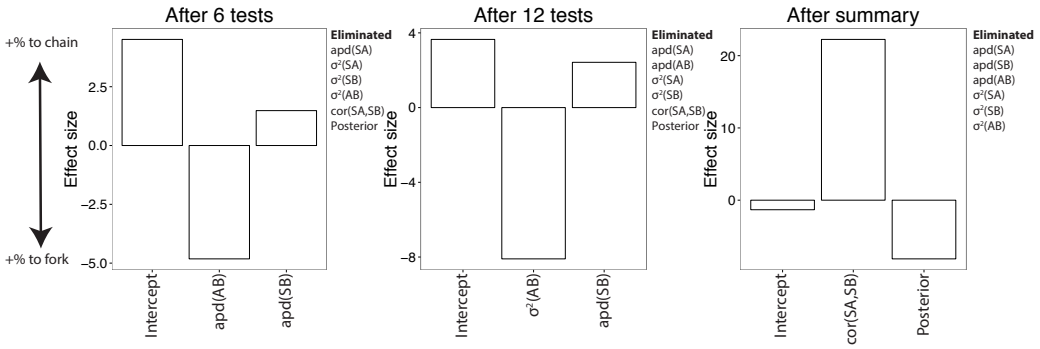

Figure 16. The models resulting from stepwise model selection to all 1040 1st, 2nd and final judgments in Experiment 3. Plots show the selected predictors' fixed effects (i.e., the β values) in descending order of t value. All predictors were z scored, and the dependent variable was centered (so that a prediction of 0 corresponded to assigning 50% to the chain and 50% to the fork). Thus, effect sizes are interpretable as differences in percentage assigned to the chain moving one standard deviation up on the independent variable.