Time in Causal Structure Learning

Neil R. Bramley
New York University

Tobias Gerstenberg
Massachusetts Institute of Technology

Ralf Mayrhofer
University of Göttingen

David A. Lagnado
University College London

A large body of research has explored how the time between two events affects judgments of causal strength between them. In this article, we extend this work in 4 experiments that 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. We focus on one-shot learning in Experiment 1. In Experiment 2, we explore how people 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 Experiments 3 and 4, 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 do not differentiate, people can still make accurate causal structure judgments on the basis of interval variability alone.

Keywords: causal learning, structure induction, time, order, Bayesian model

Many aspects of higher level cognition, including prediction, explanation, and goal-directed action, depend on representing the causal structure of the world (Sloman, 2005; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). But, how do people learn this structure? Research has predominantly focused on learning from statistical covariation between variables (Cheng, 1997; Deverett & Kemp, 2012; Gopnik, Sobel, Schulz, & Glymour, 2001; Perales & Shanks, 2007) sometimes involving 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).

In this article, we focus on the role of time in causal structure induction. To be able to predict and diagnose causality in real-world situations, we must form expectations about *how long* different relationships take to work (Griffiths & Tenenbaum, 2009; Lagnado & Sloman, 2006). For example, we learn to expect a radiator to heat up within a few minutes of turning on the boiler; to feel an effect of a pill after around half an hour; and for the TV to come on a second or two after we press "power" on the remote. Such expectations, in turn, support structure inference (Buehner & McGregor, 2006; Griffiths & Tenenbaum, 2009; Hagmayer & Waldmann, 2002; Kemp, Goodman, & Tenenbaum, 2010). If the stereo comes on shortly after we press "power" on the remote, we wonder whether we have picked up the wrong remote. If the TV comes on soon after, we might infer that both the TV and stereo are tuned to the same signal, or start looking for another cause for the stereo's activation. In general, consistency between an observed sequence of events and the predictions of different causal models provides evidence about the underlying relationships. This temporal information is unavailable to a purely contingency-based learner who only uses the statistical associations between events.

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 articles 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 (see also Spirtes, Glymour, & Scheines, 2000). The CBN framework supports probabilistic inference on structured representations, and provides a normative calculus for reasoning both about the consequences of observations as well as hypothetical (or counterfactual) interventions.

By defining a language for expressing possible causal models, the CBN 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 making inferences that are broadly normative (e.g., Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2002, 2004, 2006; Steyvers et al., 2003) but also exhibiting the signatures of inductive biases and cognitive limitations (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Dayan, & Lagnado, 2015; Bramley et al., 2015; Coenen, Rehder, & Gureckis, 2015; Mayrhofer & Waldmann, 2016; Rehder, 2014; Rottman & Hastie, 2013, 2016).

A shortcoming of the CBN framework is that Bayesian networks do not naturally encode the temporal or spatial dimensions of causal beliefs (cf. Gerstenberg & Tenenbaum, 2017). Consequently, many studies have focused on situations where information about time and space is nondiagnostic or abstracted away. When temporal cues have been pitted against statistical cues experimentally, judgments 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 often treated the observed timings of events to be diagnostic (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; White, 2006). These results suggest that we cannot tell a complete story of causal representation without accounting for the role of time. In the current article we take a novel approach: we eliminate statistical contingency information, and focus exclusively on what participants can learn about causal structure from temporal information.

The article is structured as follows. First, we review the literature on causal learning and time. After describing the learning problem we focus on, we outline the Bayesian framework and methodology we use to explore human causal learning in time. We contrast a model that attends only to event order against models that also form expectations about delays. Experiment 1 explores one-shot judgments, based on a single observation of a simple device operating through time, asking whether people's judgments are driven only by the qualitative order in which events occur, or whether they also expect similar delays across connections. Experiments 2–4 look at how people integrate evidence from multiple observations of the same device, again asking whether judgments are based only on event order, or additionally based on preference for similar causal delays. Experiments 3 and 4 test whether participants can infer the true causal structure from delay variability and correlation alone, when the order of events is held

constant but the timings of the events are either more consistent with a chain or a fork structure.

Existing Research

Temporal information is relevant for causal inference in at least two respects. First, the *temporal order* of events places hard constraints on what could have caused what. Second, within these constraints *event timings* provide additional evidence for distinguishing further between order-consistent causal structures.

Temporal Order

Because effects cannot precede their causes, the order in which events occur is an important cue to causality. In line with this, much of the human and animal learning literature is built around the notion that learners readily form associations from one event to the next (Pavlov, 1928; Skinner, 1938; Watson, 1913). Associative theories try to account for learning as an automatic pairwise association of stimuli. However, recent reanalysis of classical conditioning phenomena have suggested that learning is often better understood as involving inferences about the causal structure that is responsible for the observed events (Courville, 2006; Courville, Daw, Gordon, & Touretzky, 2003; Courville, Daw, & Touretzky, 2004, 2006; Gershman & Niv, 2012).

In the causal learning tradition, several articles have explored the role of temporal information in structure inference. Rottman and Keil (2012) investigated how people infer the causal structure of multiple variables measured at discrete time points at which variables may be subject to exogenous influences or interventions. For example, suppose you are interested in an amoeba that occasionally produces two different hormones. Suppose it is producing neither hormone at time $t - 1$. If, at time t, you stimulate the production of one of the hormones and the other hormone is also produced, this invites the inference that the first hormone causes the production of the second. Importantly, this inference is based on the fact that the second hormone's level changed state relative to the preceding time point, while pure covariational inference would treat each measurement as independent. In seven experiments, the authors found that people readily attribute causal relationships from variables influenced at time t to others whose state changed relative to $t - 1$, doing so even if a cover story strongly suggests independence (i.e., if a new amoeba is measured at each time point).

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 network and infects computers at different times. 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 displayed infection, even when covariation cues went against temporal order cues.

Event Timings

Not only the order in which events occur but also their exact timing is important for causal inference (Hagmayer & Wald-

mann, 2002). A basic associative learning result is that, as the average interval between two events increases, the associative strength between the two events decreases (Grice, 1948; Shanks & Dickinson, 1987; Wolfe, 1921). Early cognitive theories predict this effect by suggesting that the more distant two events are in time, the more costly it is to sustain the first event in working memory long enough to relate it to the second, leading to monotonic reduction in causal judgments (Ahn, Kalish, Medin, & Gelman, 1995; Einhorn & Hogarth, 1986). Buehner and May (2003) and Lagnado and Speekenbrink (2010) give a normative justification for why delays 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.

However, shorter intervals do not always lead to stronger causal inferences. Rats form food aversions even when sickness is induced hours after eating—reflecting the true time course of food poisoning (Garcia, Ervin, & Koelling, 1966). Likewise, longer fixed-interval reinforcement schedules in pigeons result in longer delays between the appearance of the reinforcer and pecking responses (Gallistel & Gibbon, 2000; Skinner, 1938). These results show that time-delayed associations are formed and can be used to guide action in animals.[1]

Humans make causal inferences that are sensitive to expectations about delays due to causal mechanisms. Seeing shorter-than-expected as well as longer-than-expected intervals leads to reduced causal strength judgments (Buehner & May, 2002; Buehner & May, 2003; Buehner & May, 2004; Greville & Buehner, 2010; Greville & Buehner, 2016; Hagmayer & Waldmann, 2002; Schlottmann, 1999). Variability in interevent intervals has usually been found to reduce causal judgments (Greville & Buehner, 2010; Greville, Cassar, Johansen, & Buehner, 2013; Lagnado & Speekenbrink, 2010), although also in one case, to increase them (Young & Nguyen, 2009). Young and Nguyen (2009) explain this increase as a consequence of experiencing occasional very short delays when there is high variability. While these studies have focused on situations in which there is a single candidate cause—effect pair, in this article, we explore the more general problem of inferring the causal structure of multiple variables based on observations of events in time.

Griffiths (2005) showed how different expectations about delay distributions allow for strong one-shot causal structure inferences. In his experiments, participants made causal judgments about "nitroX" barrels that were causally connected and exploded in different sequences. Because different causal models imply different event timings, the Bayesian model rapidly inferred the causal structure from an observed sequence of exploding barrels. Building on this work, Pacer and Griffiths (2012) model causal inference in situations where a discrete event affects the rate of occurrence of another variable in continuous time (cf. Greville & Buehner, 2007), and Pacer and Griffiths (2015) capture situations where causal influences last for some time before they gradually dissipate.

Pacer and Griffiths' (2015) approach is well-suited for capturing situations where events alter the *rate* of occurrence of other events. It does not readily apply to situations in which causes bring about their effects exactly once. In this article, we focus on situations in which the relationship between causes and effect is singular.

Modeling Causal Induction From Temporal Information

We present a modeling framework for understanding how temporal information affects causal beliefs. We first introduce our inference problem, and then lay out a Bayesian ideal-observer approach to modeling learning in this situation. We distinguish between 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

The learner's task is to 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. The learner knows that the causal relationships are generative and deterministic. Each activation of a cause component will invariably lead to the activation of its effect component(s), but the cause-effect delays are variable across instances.

For example, in the *A*-fork (see Figure 1d) 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.

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 (see Figure 1). In Experiments 3 and 4, we focus on a more restricted space of two models, with a chain from *S* to *A* to *B*, or an *S*-fork where *S* is a direct common of both *A* and *B*. In all experiments, participants see video clips that show how the components of a causal device activate over time.

Bayesian Models of Learning With Time

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 her prior probability distribution into a posterior distribution after evidence is observed. The posterior from one learning instance becomes the prior for the next. With sufficient evidence, the learner's subjective beliefs should eventually approximate the ground truth provided that the hypotheses are distinguishable[2] and the hypothesis space contains the ground truth (Chater & Oaksford, 2008; de Finetti, 1975; Lee, 2012; Savage, 1972).

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 normative yardstick for evaluating 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.

---

[1] Whether or not these behaviors rely on causal beliefs is debated (Blaisdell & Waldmann, 2012; Clayton & Dickinson, 2006).
[2] See Glymour (2001) for cases where causal structure hypotheses may not be distinguishable.
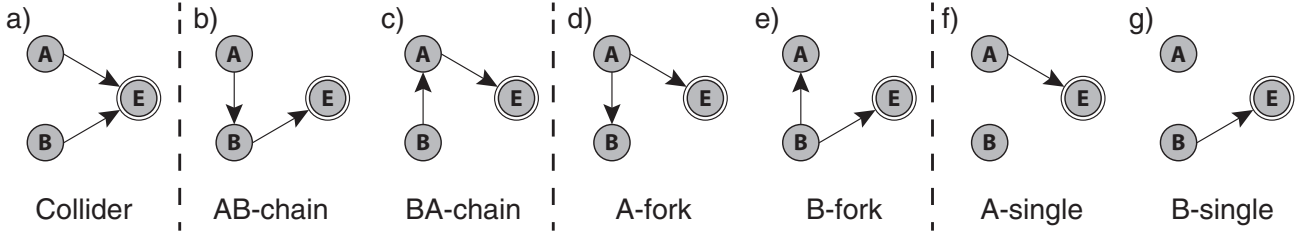
Figure 1. Possible causal structures in Experiments 1 to 3. The arrows indicate the direction of the causal relationship. Dotted lines indicate different types of structure. The collider is conjunctive—both *A* and *B* must occur for *E* to occur.

In the current context, the random variable we are interested in is the true underlying causal structure $s \in \mathcal{S}$ from 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's theorem

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

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 model based on simple likelihood functions that ignore the exact timing of events but are simply based on temporal ordering. 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 specific a priori expectations about the duration or variability of the delays.

Only Order Matters

Likelihood functions.   The order of events constrains what structures could possibly have produced the observed evidence. We capture the information contained in the temporal order of events in a simple "order 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, columns). In the following, we use the $>$ operator to denote event order. For example, $A > B > E$ means that A preceded B which preceded E. $AB > E$ 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* ($A > B > E$, $A > E > B$ and $A > BE$, see Figure 2a) but inconsistent with any pattern where either *B* or *E* precede *A*. Whether $AB > E$ or $AE > B$ are consistent with the *A*-fork depends whether one assumes causes and effects can occur simultaneously. In prior work using the same paradigm, people treated simultaneous events as inconsistent with the hypothesis that one caused the other (Bramley, Gerstenberg, & Lagnado, 2014). Thus, we assume here that simultaneous events cannot have caused each other.

When two or more structures are compatible with a particular evidence pattern, the order model favors causal structures that are consistent with fewer evidence patterns compared with structures that can produced the observed patterns as well as well as many others. For example, the *AB*-chain is only compatible with pattern ($A > B > E$ in row 2), while *A*-single is consistent with all but two patterns, and thus spreads its likelihood much more widely (cf. Figure 2). Switching focus from Figure 2's columns to its rows gives a perspective on the model's posterior predictions. For instance, upon observing a device that activates in the $A > B > E$ order (Figure 2a evidence, Pattern 2), the Order model 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) $AB > E$, the order model will rule out all structures except for the collider, *A*-single, and *B*-single. Between these remaining structures, it prefers the collider because it is consistent with fewer evidence patterns (MacKay, 1991; Myung & Pitt, 1997).

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 $p(\mathcal{S})$ to incorporate this data.

The order model only considers the qualitative ordering of the component activations, for example $\mathbf{d} = (d_1 = \{A > B > E\}, d_2 = \{AB > E\}, \ldots)$, where d_i indexes independent observations of the device. Starting from a uniform prior over the seven structures, Figure 3 shows posteriors 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), the model favors the collider after d_1, d_2, and d_3 and this preference increases with d_4. In the second example (bottom row), the model prefers the collider after d_1, d_2, and d_3 but switches upon d_4 which rules out all the structures except the A-single.

Timing Matters

Generative model.   The order model makes 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 likely to be 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 a richer representation that incorporates assumptions about how and when events are brought about.
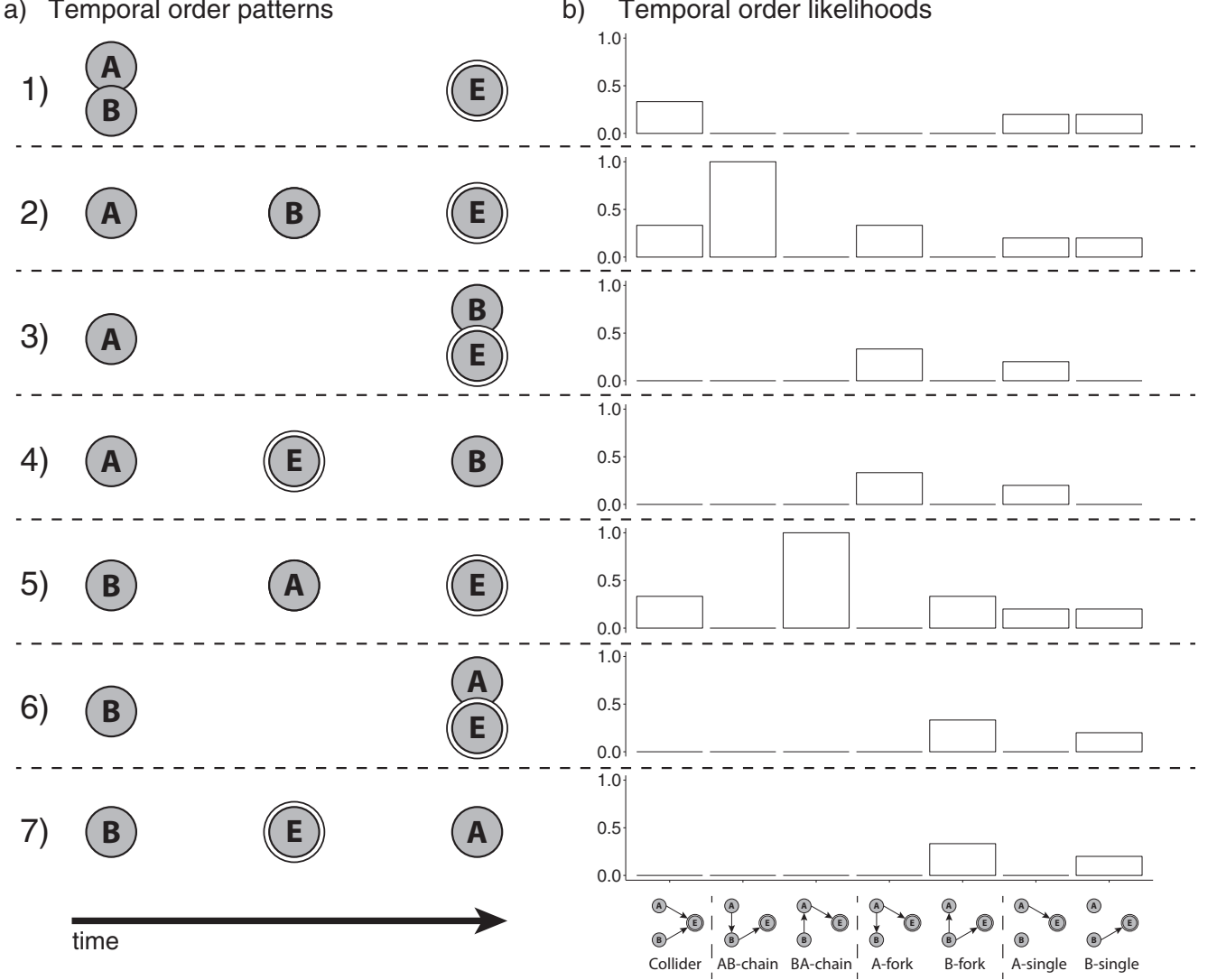
a) Temporal order patterns

b) Temporal order likelihoods

time

Collider | AB-chain | BA-chain | A-fork | B-fork | A-single | B-single

Figure 2. (a) Seven possible qualitatively different temporal patterns of three events *A*, *B*, and *E*. (b) Likelihood of each pattern type for the seven different causal structures. Each structure spreads its likelihood evenly among the qualitative temporal order patterns it could produce.

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 $d_i = \{t_A^i, t_B^i, t_E^i\}$. 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 events. We start by formalizing the timing of independent events that do not have any parents in the causal structure (such as variables *A* and *B* in the collider, *A*-single, and *B*-single). Because these events occur without being caused by other variables in the causal model, the model cannot tell us when we should expect them to occur. Thus, analogously to causal Bayes nets, in which independent causes are assumed to be statistically independent from each other (i.e., uncorrelated), we define independent causes to occur temporally independently of one another as well as independent from the (artificially determined)

beginning of a clip. We use an exponential distribution to model temporal independence. The exponential distribution is "memoryless" meaning that your expectation of when the event will occur is constant over time. If *X* is an independent event then the timing of *X* is determined by

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

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 dependence on the start time (i.e., the cause) with flexibility in the form of this dependence ranging from a strong and specific to a vague and widely distributed expectation about when the effect will occur. Gamma distributions can be defined by a shape parameter α and an expectation μ. Under the assumption that *X* causes *Y*, the timing of
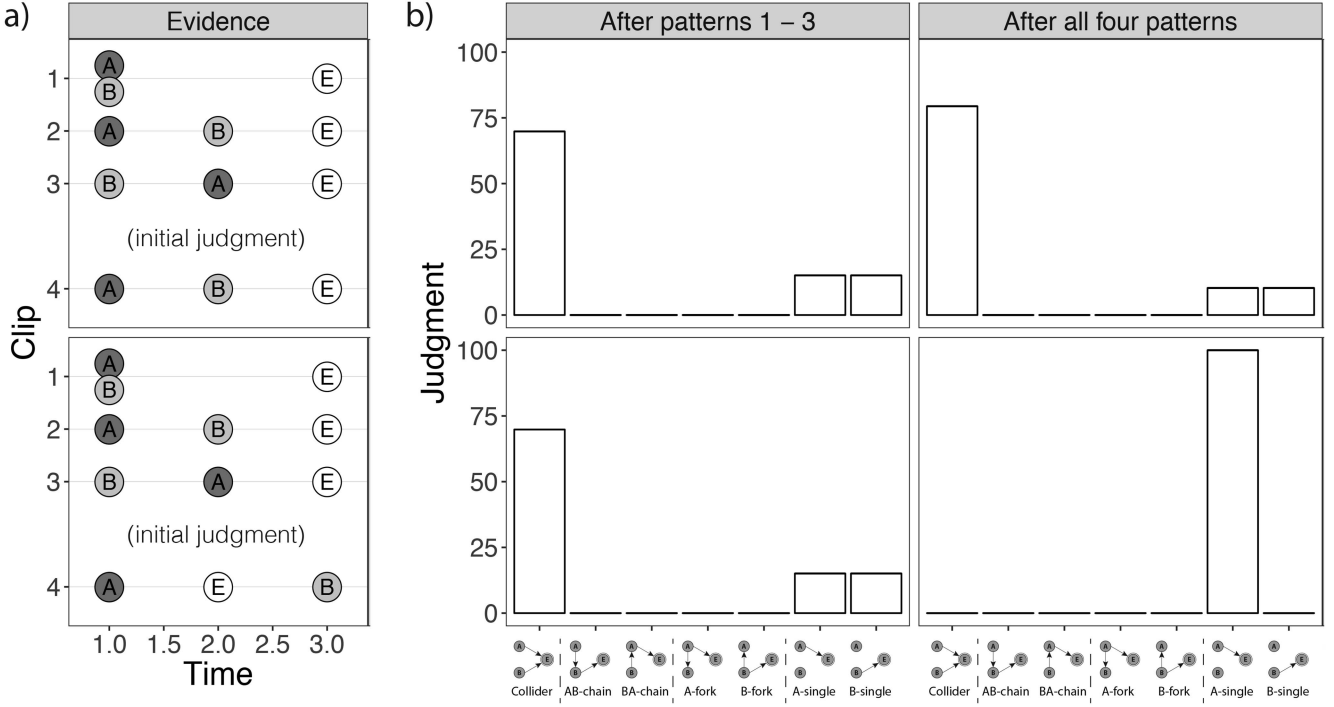
Figure 3. Two examples of predictions by the order model. (a) Sets of four time series showing staggered activation of components *A*, *B*, and *E*. (b) Model posteriors for an initial judgment after seeing Clips 1–3 (left column), and for a second judgment after having seen all four patterns (right column).

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:

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

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) delays that are variable (low α) or reliable (high α).
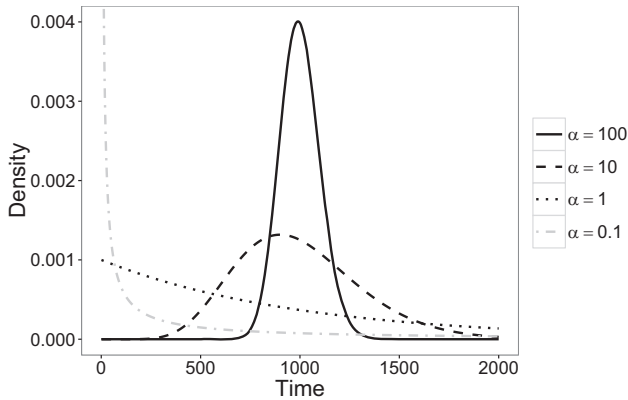

Figure 4. Example delay distributions. All have a mean of $\mu = 1,000$ ms but differ in their shape α. Temporal independence is modeled by $\alpha = 1$ (the exponential distribution). $\alpha > 1$ capture positive dependence, in which the expectation of an effect peaks some time after its cause.

As $\alpha \to \infty$, the gamma distribution becomes increasingly centered around its expected value, capturing reliable cause–effect delays (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.[3]

Colliders/common-effect structures. Within this framework, the collider (i.e., common-effect structure) requires additional thought, because it involves a joint influence of two distinct causes. There are various plausible combination functions for capturing this kind of joint influence. In the experiments reported below, we explicitly stipulate that the collider structure is *conjunctive*, meaning that the activation of *E* requires the activations of both causes *A* and *B* and, by implication, the arrival of both of their causal influences (Rehder, 2011). To model this conjunction, 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

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

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 would be modeled by simply using

---

[3] Values of $\alpha < 1$ capture what we might call "negative" dependence. This would capture a belief that, having observed a cause, one expects to see its effect either right away, or in a very long time (e.g., Figure 4, gray dot-dashed line), however we do not focus on such situations in this article.

[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.

the minimum instead of the maximum in Equation 4 (see Equations A7 and A9 in the Appendix).

Likelihood functions. The generative model laid out above provides the 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 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. This means that the same set of observed activation times is more or less likely depending on the assumed underlying causal structure and delay distributions.

Model variants. Often it may be reasonable to assume that the different connections in a causal system have the same underlying delay distribution. For instance, they might all be components of the same type. However, in other situations, we might learn to expect completely different delays for different causal relationships (e.g., burning fossil fuels produces energy quickly and environmental problems much more slowly). 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 $s \in S$. In contrast, the *independent model* (delay$_I$, Figure 5b) 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 5c) 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 article.

To see how this timing sensitivity supports causal structure inferences, let us assume that a learner observed the following order of activation: $A > B > E$. If she makes the delay$_P$ assumption that cause–effect delays for the connections in this device come from the same distribution, we would expect her 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 is most consistent with an AB-chain. However, if t_{BE} is very small this seems more consistent with the A-fork (in which $t_{A \to E}$ and $t_{B \to E}$ 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 priors on 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 shape parameter), which leads to higher likelihoods being assigned to the data point (see Figure 5a for an illustration of this point). The late occurrence of B is most consistent with an A-fork under which $t_{A \to E}$ is similar to $A \to B$, but also quite consistent with the collider where it could be the case that $t_{A \to E}$ and $t_{B \to E}$ are both short but where some time passes between the arrival of A's influence and B's influence.

Inference in the *independent* delay$_I$ model (Figure 5b) proceeds in same way, but with separate parameters for the delay distributions of the different causal connections $c \in C$; for example, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{|C|})$ 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 5b gives an illustration of this. Here, the temporal intervals t_{SB}^i for $\{i \in 1, \ldots, 12\}$ are consistently around 2s, while t_{SA}^i and t_{AB}^i are much more variable. Despite the fact that mere temporal order favors the SA-chain, we can actually 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 possible that the true structure is a chain, but the chain structure cannot explain the additional systematicity in the data whereby the variation in the $t_{S \to A}$ and $t_{A \to B}$ intervals almost perfectly cancel out.[5]

Summary

In summary, the order model operationalizes inference based purely on the qualitative ordering of observed activations. While this model is effective at ruling out inconsistent causal structures, it is limited in its ability to distinguish between structures that are consistent with the observed order of events.

The delay-based models *pooled* delay$_P$ and *independent* delay$_I$ make inferences within the space of hypotheses not-yet-ruled-out by the order model, but distribute their likelihood very differently depending on the expected rate and variability of the various interevent intervals. Assuming an uninformative prior on shape α and mean μ, the pooled delay model delay$_P$ favors whichever structures render the experienced interevent intervals the most regular across all connections and all instances (e.g., an AB-chain if t_{AB} and t_{BE} are both reliably 1 second), while the independent delay model delay$_I$ favors whichever structures imply the most regular within-edge delays on average, even if the mean causal delays differ considerably for different connections (e.g., an AB-chain if t_{AB} is a reliable 1 s and t_{BE} is a reliable 5 s).

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 Experiments 3 and 4), which were represented by circles and arranged in a triangle (see Figure 6a bottom left). During each clip, all three components activated

---

[5] 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.
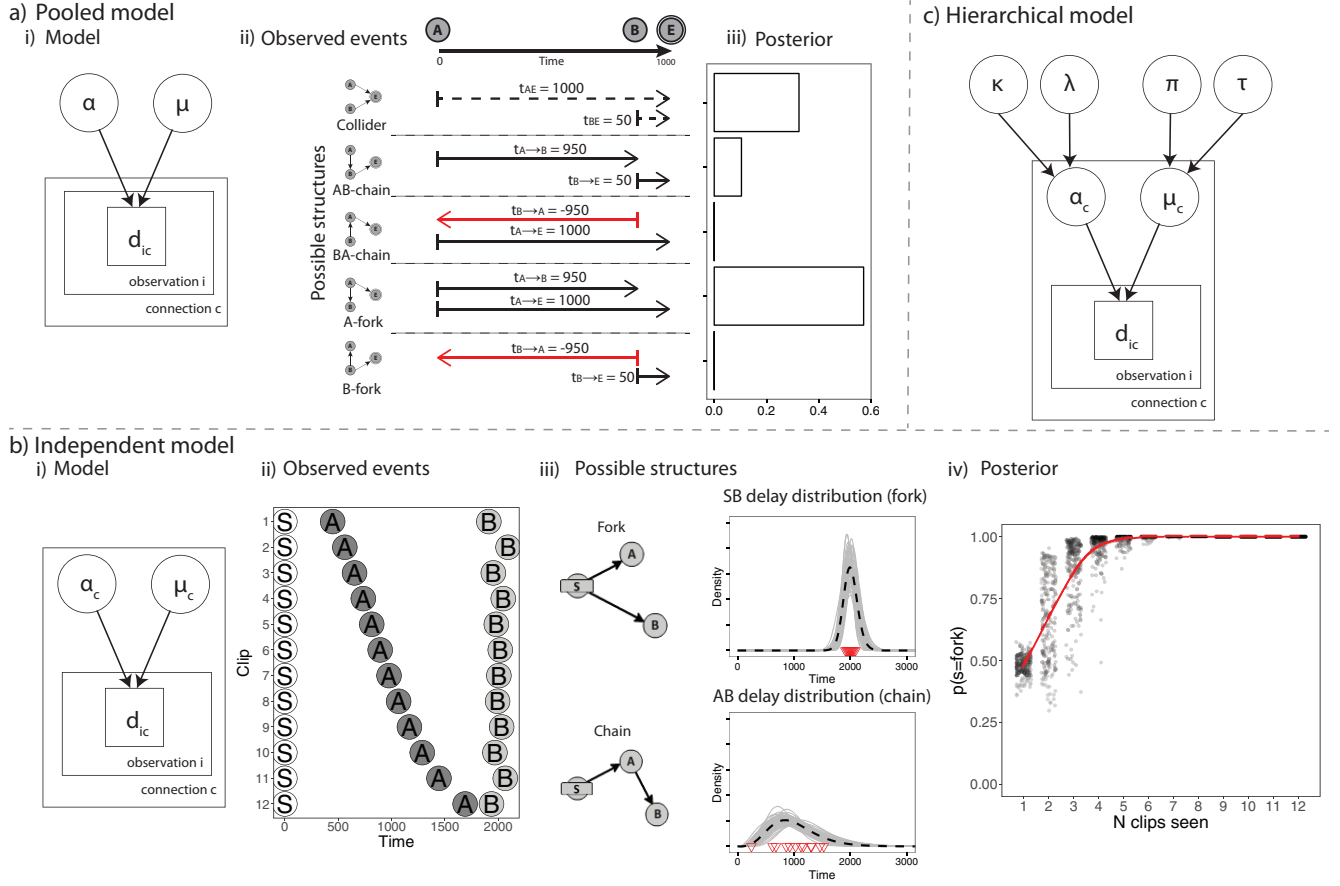
Figure 5. Delay sensitive models and predictions. (a) i. *Pooled* delay$_P$ in plate notation. ii. Going from evidence to structure inference in the pooled model. Observed event timings are mapped onto causal delays under different models. Each row shows the causal delays assuming a different structure. For the collider, dashed lines indicate that one or other causal delay may be shorter than the observed intervals. Red arrows indicate that structures can be ruled out based on order alone. iii. Posterior predictions of the delay model assuming priors of $S \sim \text{Unif}\left(\frac{1}{7}\right)$, $\alpha \sim \text{Exp}(0.1)$, and $\mu \sim \text{Exp}(0.0001)$. (b) i. *Independent* delay$_I$ model in plate notation. ii. 12 patterns of evidence ordered by increasing t_{SA}. iii. Posterior marginal inference for two possible structures. The plots show posterior observed intervals (red triangles) delay samples (gray lines) and their overall density (dotted black line). Both structures share the same $t_{S \rightarrow A}$ 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. See the online article for the color version of this figure.

by turning from white to gray (Experiments 1–2) or from white to yellow (Experiments 3 and 4). Activated components remained colored until the end of the clip. In order to separate the inference task as far as possible from domain specific knowledge of real-world causal delays, participants were not told anything about what kinds of causal processes underlie the activation of the different components but had to rely purely on the event timings.

Possible causal structures.   As discussed in the introduction, we restricted the space of possible causal structures to seven in Experiments 1 and 2 (see Figure 1) and two in Experiments 3 and 4. 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 Experiments 3 and 4 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 6a top). We can then directly compare participants' distributions over the structures with the predicted posterior distributions based on our different models.
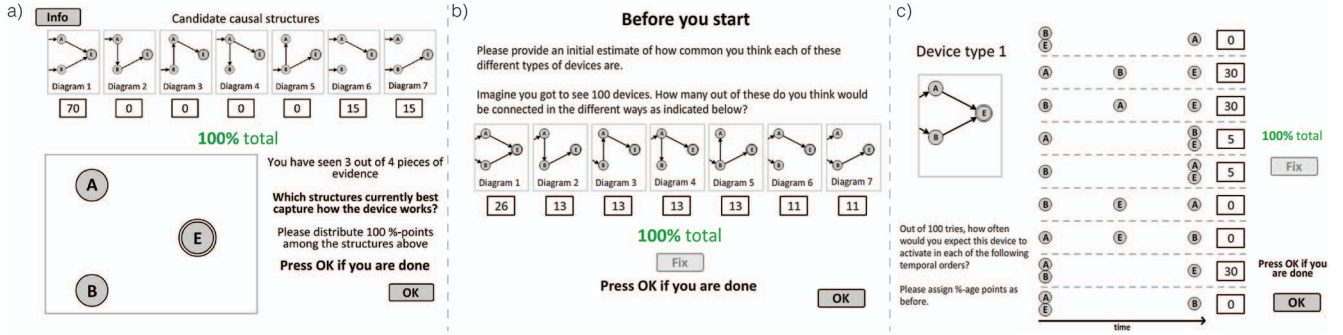
Figure 6. (a) Experiments 1–2 main judgment interface. (b) Experiment 2: Eliciting priors before main task. (c) Experiment 2: Eliciting likelihoods after main task. See the online article for the color version of this figure.

Experiment 1: One-Shot Inferences

Often in everyday life we must make snap judgments about causality based on a single observation. In Experiment 1 we explored this kind of "one-shot" inference. We asked participants to make judgments about causal devices after watching a single clip and replaying it several times. We were interested simply in the extent and nature of judgment differences depending on the relative timing of the activations. We varied the timing and order of the activation of the three components systematically across problems. In particular, we were interested whether, in the absence of specific delay expectations, judgments would reflect a delay$_P$ preference for models under which the cause–effect delays are similar between components. If this was the case, we would expect participants to assign more points to, for instance, the collider if B occurs very early as in Clip 2 shown in Figure 7a, and the fork if B occurs very late as in Clip 6.

Method

Participants and materials. Thirty-one participants (18 female, M_{age} = 36.8, SD_{age} = 11.9), recruited from Amazon Mechanical Turk[6] took part in Experiment 1. The task took 15 min (SD = 8.7) on average and participants were paid at a rate of \$6 per hour. The task interface was programmed in Adobe Flash 5.5.[7] Demos of all experiments, along with source code and data is available at https://github.com/neilbramley/time_cause.[8]

Stimuli and model predictions. Participants made judgments about nine devices in total. For each device they saw evidence in the form of a single, replayable video clip. All clips began with a 500-ms interval after which the first component(s) activated. The clip then lasted another 1,000 ms whereupon the final component(s) activated. We chose a range of clips that varied where the activation of B fell between A and E (see Figure 7a, Clips 1–7), and also included two clips in which E occurred earlier than B (Clips 8 and 9). Participants were told that the parentless components in the device activate due to background causes. We obtained model predictions by computing the posterior for $p(\mathcal{S}|\mathbf{d})$ for order and delay$_P$, assuming learners began each problem with a uniform prior across structures (see Appendix).[9] We compare these against a Baseline model that simply assigns an equal probability to all models.

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 favors the A-fork 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.

Procedure. In the instructions, participants were familiarized with the seven causal structure diagrams, and the response format. Participants then completed the nine problems in random order. Components A and B were counterbalanced such that on approximately half of the problems their roles were reversed (e.g., B would occur at the start rather than A) with 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 seven 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 6a while for the other half it was reversed.

Results

We compared aggregated participant judgments to our models by correlating the assignments across the seven structural hypotheses for each judgment and then averaging these correlations across the nine devices using the Fisher's z transformation (Corey, Dunlap, & Burke, 1998). The average assignments of percentage points over the seven structures were best correlated with the

---

[6] 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).

[7] Flash has been shown to be a reliable way of running time-sensitive experiments online (Reimers & Stewart, 2015).

[8] All experiments were approved by UCL Research Ethics Committee under protocol CPB/2014/006.

[9] Note that delay$_I$ requires repeated evidence to form preferences about causal structures.
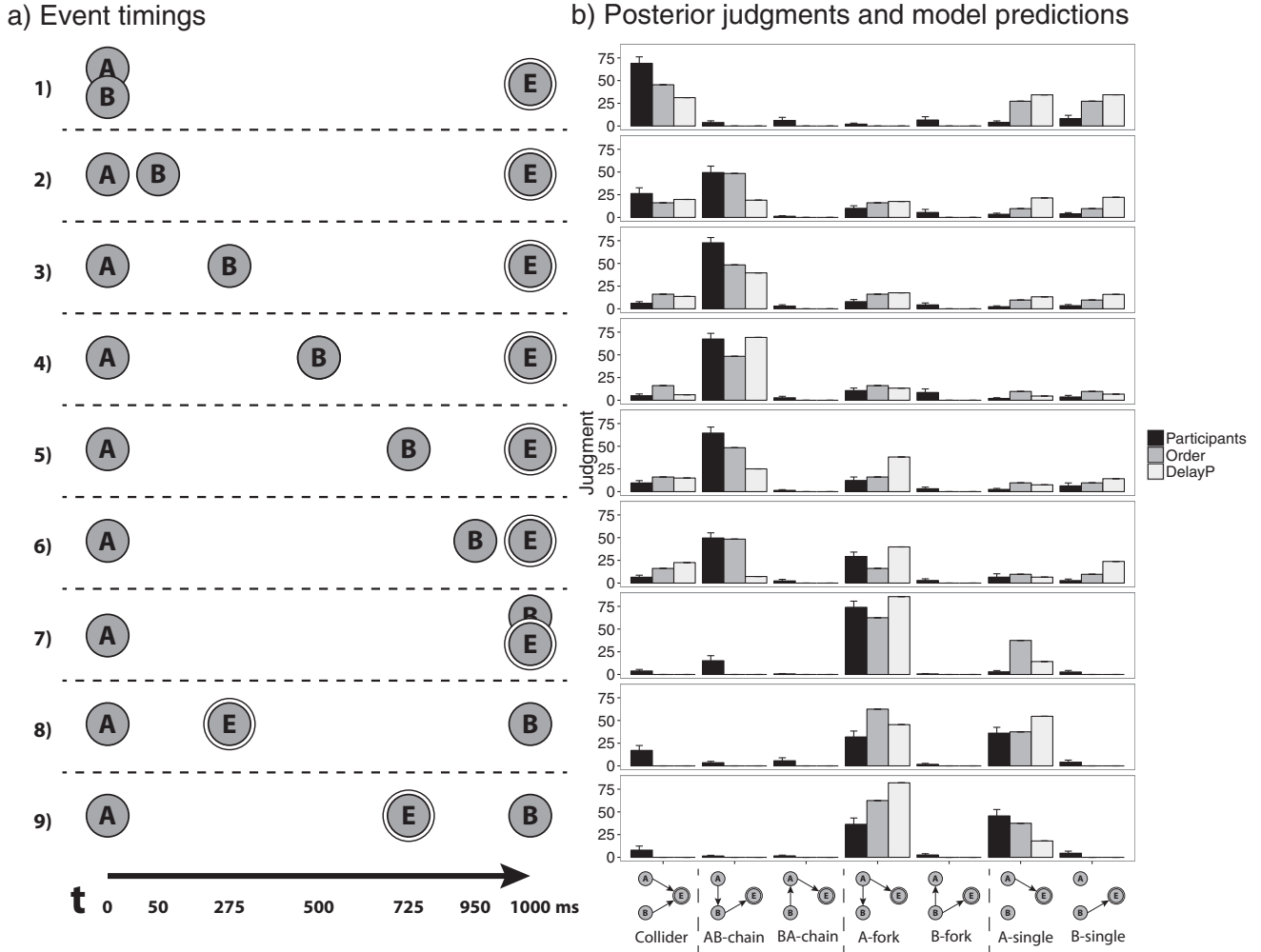
Figure 7. (a) The timeline for each clip type in Experiment 1. (b) Participants' averaged judgments after viewing each clip (black bars) and predictions by the models (gray bars). Error bars ±1 standard error.

order model, and to a lesser degree with delay$_P$ (see Figure 7b and Table 1).

Overall there was relatively little sensitivity to the exact timing at which B occurred. If we compare Patterns 2 to 6 in Figure 7b, we see that the AB-chain was the modal response across early to late occurrence of B consistent with order predictions. Notwithstanding, there was evidence of modest time sensitivity. Figure 8 shows participants' probability assignments to the collider, AB-chain, and A-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 A-fork. This timing sensitivity is revealed by fitting repeated measures ANOVAs to the points assigned to the collider, the AB-chain, and the A-fork across Clips 2–6. The three structures' assignments vary across these clips (collider: $F(4, 120) = 7.1$, $p < .005$; AB-chain: $F(4, 120) = 5.1$, $p = .002$; A-fork: $F(4, 120) = 10.0, 27$, $p < .001$, with Greenhouse-Geiser corrected p values) while the order model does not distinguish 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 4) under which Clip 6 is consistent with equal (e.g., 50 ms) causal delays $t_{A\rightarrow E}$ and $t_{B\rightarrow E}$ with A's influence arriving earlier and "waiting" for B's, while not a perfect match to either chain

Table 1
Experiment 1: Order and Delay Models Compared With Judgments

Model	$M\ r$	$SD\ r$	$M\ r_s$	$SD\ r_s$	Mode match	RMSE	N
Baseline	0	0	0	0	.14	20.1	0
Order	.90	.29	.76	.26	.78	11.1	22
Delay$_P$	.80	.67	.64	.39	.44	15.9	9

Note. M and $SD\ r$ = mean and standard deviation of Pearson's r correlation between model and participants' assignments to structures within each device. M and $SD\ r_s$ = same for Spearman's rank correlations; Mode match = proportion of problems where participants' modal choice matched model's; RMSE = root mean squared error (%); N = number of individuals best correlated with model.
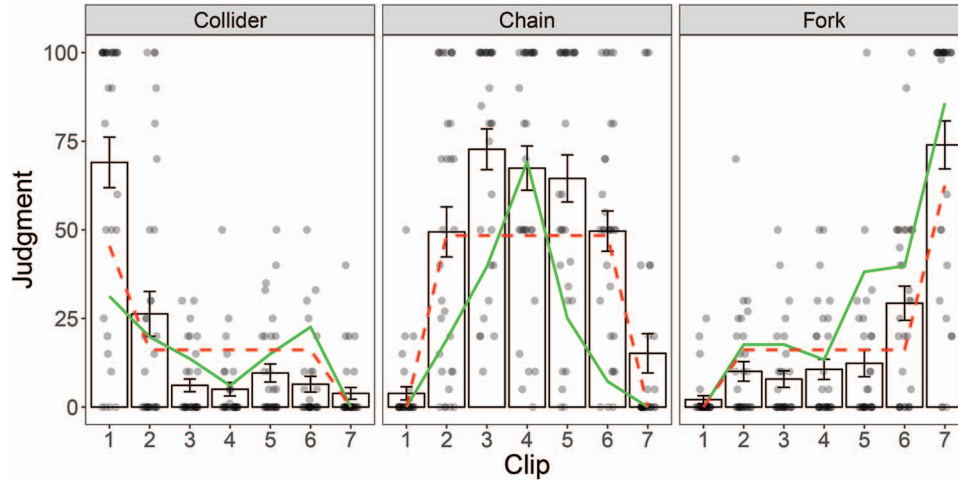
Figure 8. Comparison of probability assignments to fork, chain, and collider structures for Clips 1–7 (cf. Figure 7), in which *B* appears at 0 ms, 50 ms, 275 ms, 500 ms, 725 ms, 950. and 1,000 ms after *A*, with *E* always occurring at 1,000 ms. Bars show participant means with standard errors with points, showing individual judgments, jittered on the *x*-axis for visibility. Results in text concern the five clips with staggered activations 2–6. Red dashed line denotes order model predictions, green full line delay$_P$ predictions. See the online article for the color version of this figure.

or fork. As expected, *AB*-chain judgments peaked when t_{AB} and t_{BE} are the same (Clip 4) and *A*-fork when t_{AB} and t_{AE} are the same (Clip 7).

Discussion

The results of Experiment 1 show that participants' one-shot structure judgments can largely be explained by a simple model that only uses event order. The order model's predictions were not perfect though, underestimating participants' strength of preference for the collider in Clip 1, chain in Clips 3–5 and fork in Clip 7, and assigning more weight overall to the *A*- and *B*-singles.

One possible explanation is that participants found some structures more likely a priori than others. Alternatively, 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 independently at different times. Another, more indirect, explanation could be that participants figured that in the collider, *A*, and *B* produce *E* only if they occur roughly simultaneously. Because we stipulate that we focus on cases where all the events do occur, this might be seen as implying that any trials where the true device is a collider must be ones where *A* and *B* did in fact occur at a similar time. Finally, for the fork structure, people might assume that the two effects share an intermediary delay mechanism, such that the effects will tend to occur simultaneously even if their delay might vary if the device was tested multiple times (Park & Sloman, 2013).

Participants' judgments shifted over Clips 2 to 6 as predicted by the delay$_P$ model. This is evidence for some default assumptions about timing. However, it was not sufficient to alter many participants' modal judgments away from those predicted by the order

model. Figure 8 shows an inverted *U* pattern for the *AB*-chain across Clips 2 to 6, rather than the inverted *V* shaped curve predicted by the delay$_P$ model. A possible 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 (Addyman, French, & Thomas, 2016). Generally, participants exhibited a robust preference for a chain structure whenever activations occurred sequentially.

In Experiment 1, participants had little evidence to go on. Observing a device in action only once, does not reveal its full range and variability in behavior and limits the scope for forming expectations about delays.[10] Thus, to further 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 must integrate this evidence to narrow in on the true causal structure.

Experiment 2: Integrating Evidence

In this experiment, we explored how people integrate evidence from multiple observations of the same device. We constructed sets of evidence in which we varied the order in which components activated. Participants saw several different clips of each device, made an initial judgment, and were then asked to update their judgment after an additional clip. This procedure allowed us to explore how learners revised their beliefs as they received more evidence. Additionally, by providing multiple exposures to the same set of connections, we were able to test whether participants' judgments reflect a preference for consistent cause-effect delays.

---

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

Participants in Experiment 1 made larger assignments to the collider and smaller to the A- and B-single structures that our order and delay$_P$ models predicted. We hypothesized that this could be partly due to different a priori assumptions about which kinds of structures are more probable. Some participants' might have thought that, because the collider was the only unique structure type among the response options (see Figure 1), it would be encountered more frequently, or that the single structures, with their unconnected components, were somehow degenerate devices, and so less likely to be encountered. Another possibility is that, while participants may rely predominantly on temporal order, they might still judge the likelihood of order patterns differently than simply dividing evenly across order-consistent patterns as we initially assumed. In particular they might think qualitative patterns that imply equal delays (such as $A > BE$) under the A-fork) are more likely than those that imply unequal delays (such as $A > B > E$ under the A-fork). We tested both of these hypotheses directly in Experiment 2 by eliciting judgments of *prior* probability for the different device types—that is, $P(\mathcal{S})$—and judgments of the *likelihood* that different device types would produce the various order patterns—that is, $P(\mathbf{d}|\mathcal{S})$. This allowed us to assess the extent to which participants' integration of evidence was normative with respect to their assumptions and expectations about how different devices would behave.

Method

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

Stimuli and model predictions. We created evidence sets for eight different devices. For each device, participants watched four clips. We depict the qualitative order of the events in each clip for each device in Figure 9a. Participants were asked to provide a posterior probability judgment after they had seen the first three clips, and another after having seen the fourth clip. We selected sets of clips such that, for Devices 1–3, little or no shift in the most favored device was predicted by the order model, while for Devices 4–8, the order model predicted a large shift belief between the first judgment (after the first three clips) and the second judgment (after the fourth clip).

For example, Devices 1 and 4 correspond to Figure 3 top and bottom. Thus, for Device 4, participants first saw a clip in which components activated in the order: $AB > E$, then one where they activated $A > B > E$ and one where they activated $B > A > E$ resulting in a strong prediction by the order model in favor of the collider. Finally, participants saw a fourth clip where components activated $A > E > B$, which is incompatible with a (conjunctive) collider, meaning that the order model predicts a dramatic shift to A-single—the only remaining structure that is consistent with all four patterns. We only used sets of patterns that did not lead any of the considered models to rule out all the causal structures.

Because we elicited individuals' priors and order-based likelihoods, we can compute the **I**ndividual order model's posterior predictions $P^{IV}(\mathcal{S}|\mathbf{d})$ using participant's subjective likelihood functions $P^{IV}(\mathbf{d}|\mathcal{S})$ and prior $P^{IV}(\mathcal{S})$.

In the experiment, we drew intervals between components independently for each connection. Because each participant might have a different prior, we also compute the delay$_P$ model's predictions separately for each participant.

Procedure. After reading the instructions, participants had to successfully answer comprehension check questions to proceed. At this point there was 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 6b). Then participants completed a *posterior judgment phase*, as in Experiment 1. Here participants made judgments about the structure of eight devices, one after watching three clips and then a second judgment after the fourth clip in each case (see Figure 9). At judgment time, they were provided with a qualitative visual summary of the clips they had seen for that device. The order in which the devices were presented was randomized between participants. However, the order of clips for each device was always as shown in Figure 9a. We varied the interval between each activation, drawing each from a uniform distribution between 200 ms and 1,200 ms to assess the role of delays at the individual level. The clips used in the experiment varied in total length between 1,189 ms and 3,094 ms depending on these intervals and whether there were three staggered component activation events or only two. We again counterbalanced two presentation orders of the seven structure hypotheses shown at the top of the screen between participants (Figure 6b).

Finally, participants completed a *likelihood judgment phase*. In this phase, participants made seven additional percentage allocations, one for each causal structure. We elicited these at the end so that a desire for consistency could not affect participants' posterior inferences (Koriat, 2012). However, we note that a preference for self consistency could mean that these likelihood judgments were influenced by recall of posterior judgments. For each allocation, participants 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 percentage points across the different temporal order patterns (see Figure 6c). 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.

Results

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

Prior judgment phase. We used clustering to identify whether participants break down into meaningful groups in terms of their judgments of prior probability across the seven structures. We used a finite mixture model to find the BIC optimal clustering. This analysis found seven clusters, respectively, containing 18, 11, four, three, two, one, and one participant.[11] Eighteen participants assigned roughly equal weight to all seven response options (see Figure 10a). Eleven assigned approximately double to the collider compared with the rest of the structures. The remaining 11 participants made heterogeneous responses.

---

[11] We used a Gaussian finite mixture model using R's mclust package (Fraley, Raftery, & Scrucca, 2012).
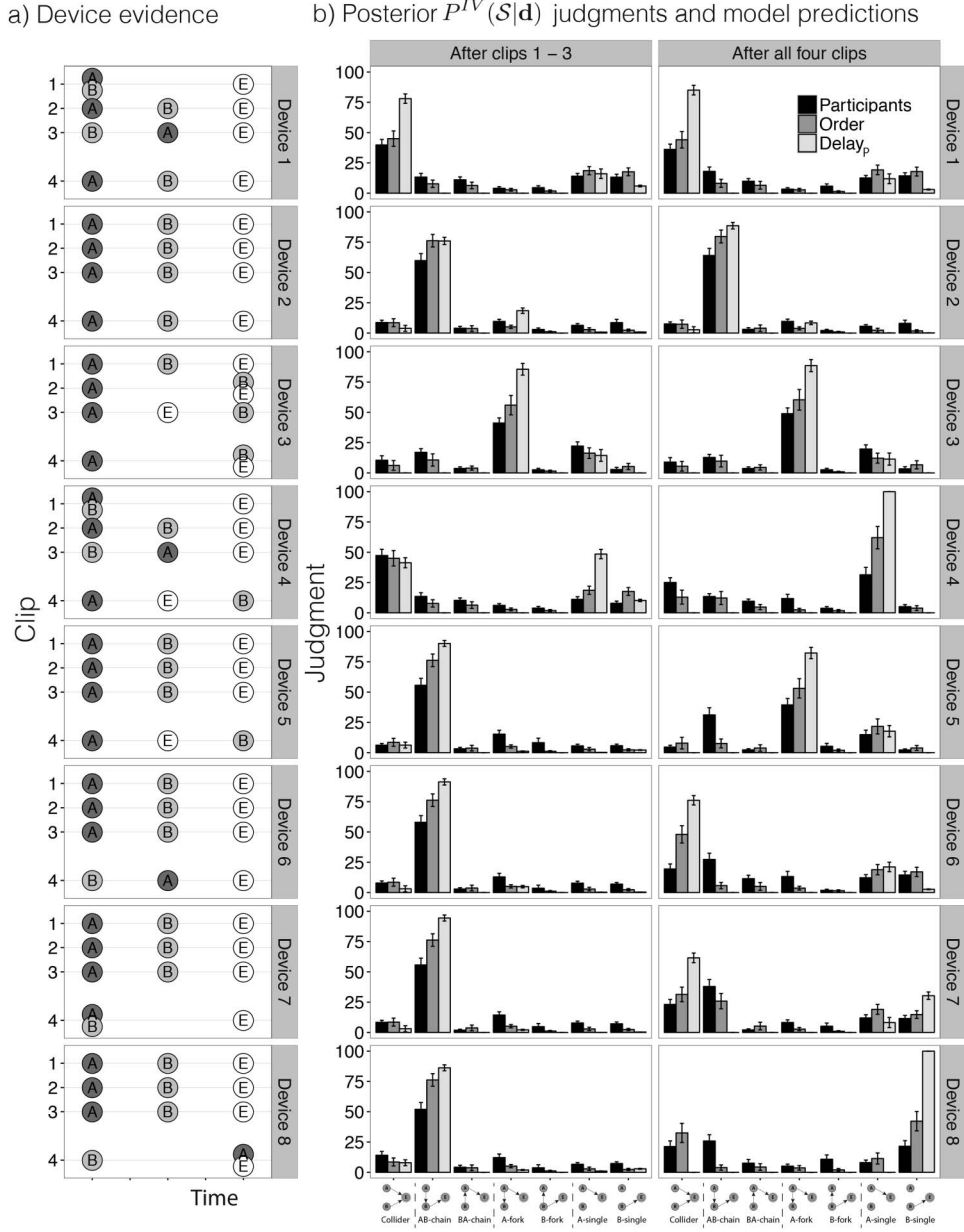
Figure 9. Experiment 2 devices and posterior judgments. (a) Devices and qualitative order of activations for each. (b) Participants' posterior judgments (black bars) compared with a model based on individually elicited priors and order-based likelihoods Order (dark gray bars) and individual priors and exact timings delay$_P$ (light gray bars). Left hand column, judgments after viewing three clips, right hand column judgments after all four clips. Error bars ±1 standard error. Model prediction bars omit cases in which participants' chosen likelihoods and priors led to all hypotheses being ruled out.

Likelihood judgment phase. Participants' likelihood judgments across the nine temporal order patterns for each of the seven structure hypotheses are shown in Figure 10b and 10c. Both the order model as well as the delay$_P$ model accurately capture participants' likelihood judgments (see Table 2). For the delay$_P$ model we assumed the same parametrization as in Experiment 1, and encoded the timings implied by the depictions of the order patterns (e.g., Figure 10b).[12] However, as

Figure 10 reveals, only the delay$_P$-derived likelihoods capture the fact that participants assigned more probability to the patterns implying reliable delays—more to $AB > E$ than $A > B >$

---

[12] Specifically assuming they represented a total interval of 1,400 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.

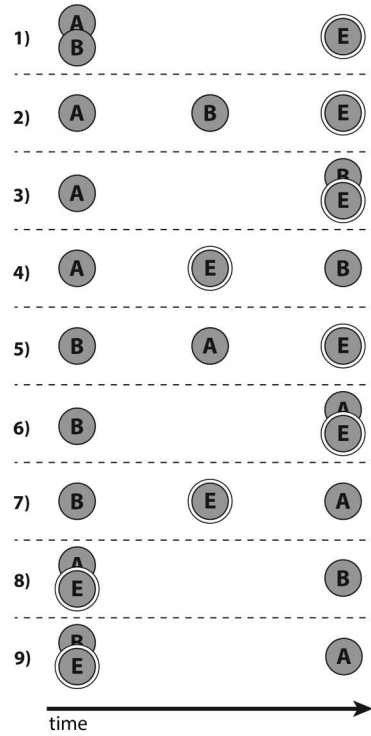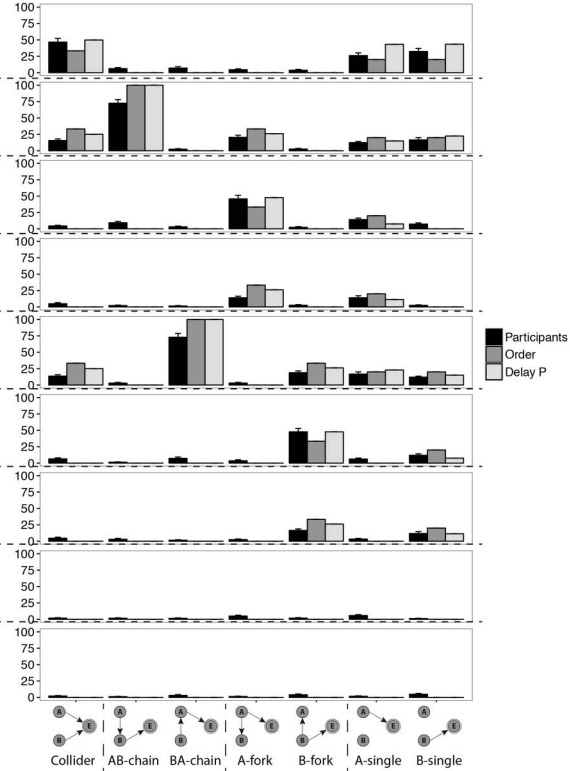a) Clustered prior $P^{IV}(\mathcal{S})$ judgments
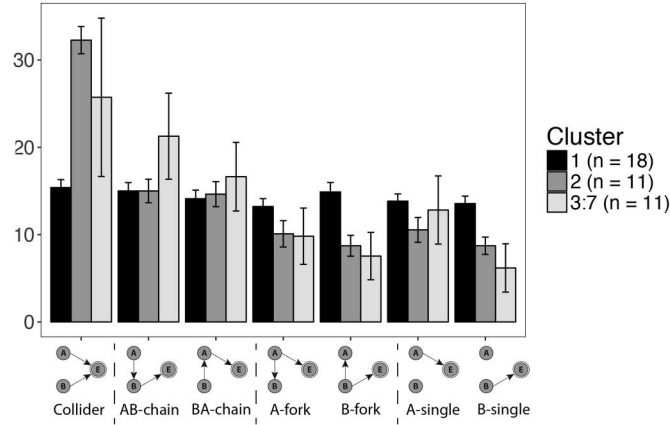
b) Qualitative evidence patterns   c) Likelihood $P^{IV}(\mathbf{d}|\mathcal{S})$ judgments and predictions

Figure 10. Elicited prior and likelihood judgments. (a) Prior probability judgments, clustered as detailed in text. (b) Nine temporal order patterns. (c) Participants judgments of the likelihood of each pattern compared with those of order and delay models as detailed in text. Error bars indicate ± 1 standard error.

E or $A > E > B$ for the collider, and more to $A > BE$ than $A > B > E$ or $A > E > B$ for the A-fork, and similarly for the B-fork—while the order model spreads its likelihood equally across these cases.

Posterior judgment phase.   Because each participant experienced different timings, we compared individual's posterior judgments rather than aggregated judgments with individually computed versions of our order and delay$_P$ models (Figure 9b). For each participant, we computed order based on their personal

prior $P^{IV}(\mathcal{S})$ and likelihood $P^{IV}(\mathbf{d}|\mathcal{S})$ judgments. Similarly, we compute individual delay$_P$ predictions based on participants' personal priors plus the sensitivity to the precise event timings they experienced. We find a closer by-problem correlation between the time-sensitive delay$_P$ model's predictions and participants' judgments with an average correlation of .78 (SD = .58) compared with .70 (SD = .59) for the order model. Twenty-four participants were better correlated with the delay$_P$ model and 15 with the order model. One participants was not

Table 2

Experiment 2: Likelihood Judgment Model Fits

Model	$M\ r$	$SD\ r$	$M\ r_s$	$SD\ r_s$	Mode match	RMSE	N
Baseline	0	0	0	0	.11	15.4	2
Order	.92	.30	.78	.30	.71	8.9	12
Delay$_P$	.98	.39	.81	.39	1.00	7.3	26

Note. r = average Pearson's r correlation between average likelihood judgments for patterns under different device types, and likelihoods derived from models; 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.

correlated with either model.[13] This suggests some influence of experienced delays over and above the order of the events for many participants. However, the idiosyncratic differences in what the delay$_P$ model predicts from these delays prohibits focused analysis of where participants did and did not make additional use of temporal information.

The fact that many individuals' judgments were better correlated with the delay$_P$ model than the order model invites the question of whether participants were generally influenced by a mixture of order and delay information, or whether some individuals focused on order and others on delays. We assessed this by fitting individual regression models, each predicting a single participant's judgments with both their personalized delay$_P$ model and order model predictions—using their priors and, for order, their likelihoods—and comparing the strength of the contribution of the two factors for each participant. Twelve participants had a significant positive relationship only with the order model, eight only with the delay$_P$ model, while there were 11 for whom both the order model and the delay$_P$ model had a significant positive relationship with judgment patterns, with the remaining nine unrelated to either measure (four) or negatively related to one or other measure (five). This analysis suggests substantial heterogeneity in participants' inferences, with many individuals focusing primarily on order information, some primarily on timing, and others on both order and timing.

Discussion

In Experiment 2, we attained a clearer picture of the sources of variability in people's causal structure inferences. Eleven participants considered the collider structure to be significantly more probable a priori than the rest of the structures. One possible explanation for this is that these participants conceived of the options in terms of four *types* of structure: collider, chain, fork, and single, subdividing across the specific cases within each type. This would explain their putting more prior weight on the collider, because it is the only structure option in its class. This is sensible from a taxonomic perspective (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008) suggesting that participants are simply registering a uniform prior over what *type* of structure they expect to see.

Participants reported finding patterns that implied equal delays across all components more likely than unequal delays. This finding was predicted by the delay$_P$ model which assigns more likelihood to these patterns. Combining personal priors and qualitative

likelihoods resulted in a tight qualitative fit with participants' posterior judgments. Furthermore, many participants showed hints of additional sensitivity to timing, with posterior predictions based on their individually experienced event timings (delay^{IV}) contributing significantly to explaining their judgments over and above what was accounted for by their order preferences. In general, these results suggest that an expectation of regularity feeds into structure judgments even in the absence of domain knowledge.

Interestingly, despite distributing likelihoods in a way that suggested they preferred equal delays across devices' components, a substantial number of participants' posterior judgments were better described by the order model than the delay$_P$ model. One explanation for this is that the design of the experiment 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 may have encouraged some participants to focus on order.[14]

In sum, Experiments 1 and 2 show that participants make sensible judgments about structure based on event order, and to some extent, event timing even in the absence of prior expectations about the nature of cause–effect delays. To look more closely at the role of timing, we now turn our focus to a more constrained situation where order is nondiagnostic and the only available information comes from the variability and correlation in event timings.

Experiment 3: Learning From Timing Variability Alone

In this experiment, we focused on causal inference in situations in which the qualitative order of events was held constant, but the temporal delays between events varied. 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. We systematically varied the mean and variability of the interevent timings such that they were more consistent with having been generated by either a chain or a fork under the delay$_I$.

As we will show in the Model Predictions section below, for a fork and chain to be distinguished based on timing alone, there must be variability in the delay leading to the intermediate event— here $t_{S \rightarrow A}$—and there must be not too much variability in the other delay—$t_{S \rightarrow B}$ or $t_{A \rightarrow B}$. This produces correlation between the timing of A and the timing of B in the case that the structure is in fact a chain. If there is variability in the other delay, this correlation will be weaker but still present. This correlation fits with the Markov condition (Pearl, 1988): A and B should be dependent in the chain, even controlling for S. S is held constant because subjects initiate the system by activating S. In contrast, for $A \leftarrow S \rightarrow B$, A and B should be independent conditional on S. Again, because subjects

---

[13] For 14% of first and 27% of second judgments, all structure hypotheses were ruled out based on combining individuals' priors and likelihoods. To allow comparison we simply had the models 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] However, Bramley et al. (2014) included an experiment similar to the posterior judgment phase of this experiment but varied whether participants saw a qualitative, quantitative, or no visual summary. Results were highly correlated across conditions .94 and no systematic differences in judgment patterns was observed.

intervene on S, and timings are relative to S, S is being controlled for.

We hypothesized that participants would be sensitive to these differences and 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 $S > A > B$ 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 qualitative differences in event order. Thus, we predicted that participants would be more uncertain in their posterior judgments, requiring more evidence even accounting for the narrower response set. To assess how well participants detect and track timing variability, we first elicited judgments based on simply experiencing the timings. Afterward, we provided participants with summaries of the trials detailing all the timings visually, and allowed them to update their judgments. By providing participants with summaries, we eliminated any potential memory effects, or effects resulting from perceptual noise associated with encoding the timings (Addyman et al., 2016). We expected participants' preference for one of the two structures to become stronger and more normative after having seen the visual summary.

Method

Participants and materials. One-hundred and four 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 sample size was determined by the number of students on the course, who all participated in the experiment on the same day. The task took M = 23.0 min (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 assigns more likelihood to the evidence (marginalizing over its unknown parameters).

To explore this we chose a set of devices that exhibited a mixture of reliable and unreliable causal delays (see Figure 11a). We created two chain devices (*chain1* and *chain2*) and two fork devices (*fork1* and *fork2*) by augmenting each connection with a delay distribution (see Figure 11a). All four devices shared an $S \rightarrow A$ connection with delay parameters α = 5, μ = 1,000 ms. This meant that A would occur an average of 1,000 ms seconds after S but with considerable variability (SD = 447 ms). We then chose distributions for $A \rightarrow B$ for the chains and $S \rightarrow B$ for the forks such that t_{SB} was 2,000 ms 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 interval between A and B (α = 1000, μ = 1000 ms). In *chain2* $t_{A \rightarrow B}$ had as much variability as $t_{S \rightarrow B}$. In *fork1*, the $S \rightarrow B$ connection had a near-constant 2-s delay (α = 1000, μ = 2000 ms) while in *fork2* the delay was variable (α = 10, μ = 2,000 ms).

We used these four generative devices to select sets of 12 clips used as evidence. We chose a representative sample of clips by taking 12 equally spaced quantiles from each distribution. To ensure that the delay draws for $t_{S \rightarrow B}$ (or $t_{A \rightarrow B}$ for the forks) were independent of those for $t_{S \rightarrow A}$, they were paired in counterbalanced order. The resulting sets of evidence are depicted in ascending order of t_{SA} in Figure 11b. Finally, we included a variant of *fork2*, named *fork2rev*, which featured a single order reversal trial. This allowed us to compare the respective influence of order and timing cues on participants' judgments.

Model predictions. We used the delay$_I$ model to obtain a posterior joint distribution over the true structure (i.e., fork or chain) and its associated parameters.[15] 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 11e 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 twelfth clip, all participants have seen the same evidence so the predictions converge.[16]

Figure 11e 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 11c 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 more mass to all the data points which are less spread out and more evenly distributed under the true than the alternative structure. Under the noisier *chain2* and *fork2* evidence, the model forms infer the correct structure 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. 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, because 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

---

[15] 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.

[16] We used MCMC to estimate these posteriors without specifying any prior on delay parameters. In the Appendix we compare these with Simple Monte Carlo sampling predictions under a variety of priors. This allows us to assess the impact of prior choice for α and λ in Experiments 1 and 2.
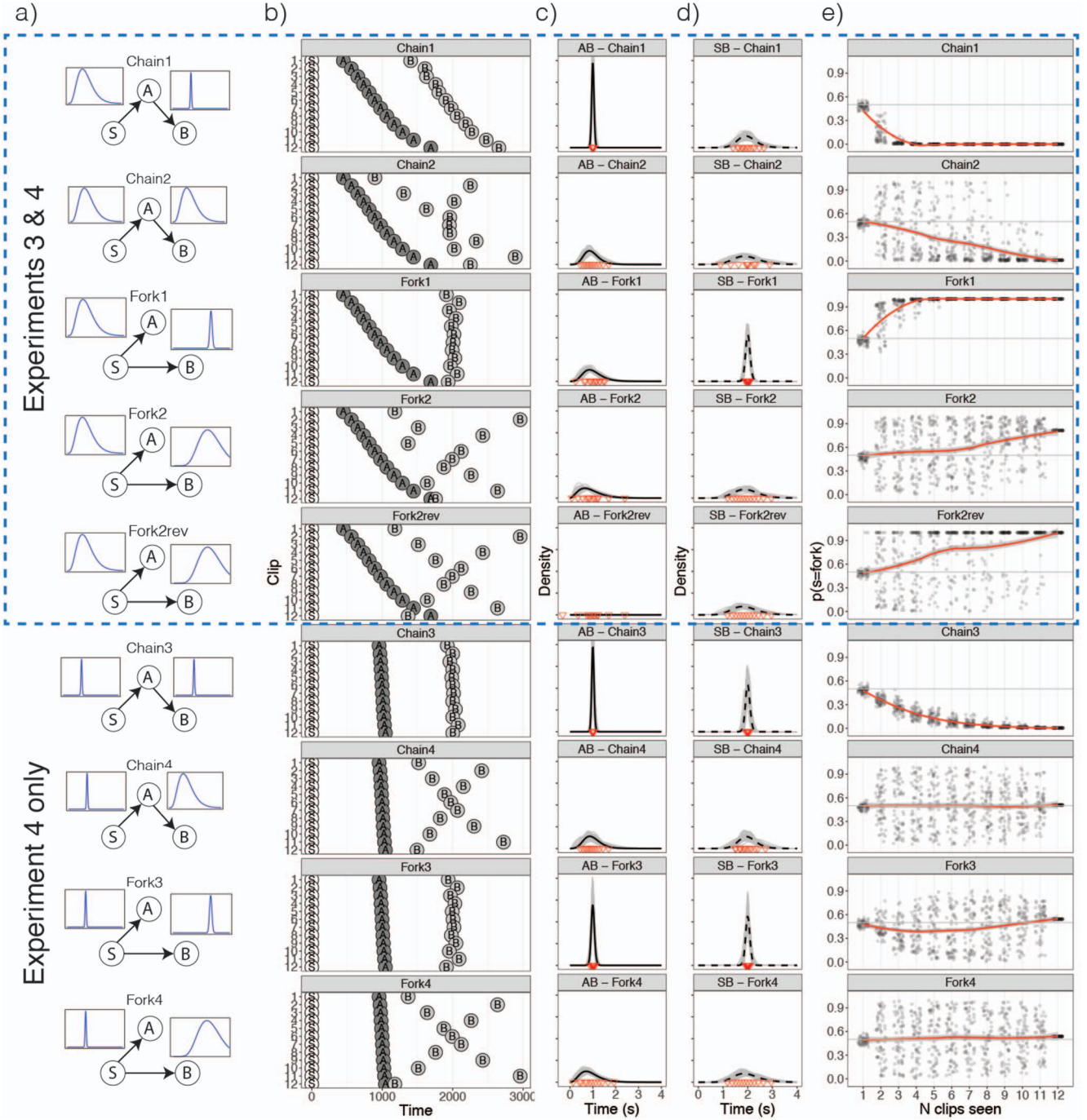
Figure 11. Experiment 3 and 4: Stimuli and model predictions. (a) Graphical representation of the different device types. (b) Plot showing the 12 patterns generated for each device. (c) Red triangles: t_{AB} for Patterns 1 to 12. Gray lines: $P(G_{A \to B}|\mathbf{d})$ for a posterior sample of αs and 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. (c) Posteriors $P(s = fork|\mathbf{d})$ for progressively more evidence. Individual dots represent the samples of evidence seen by participants. The red lines represent smoothed average (using the general linear additive model with integrated smoothness estimation gam from R's mgcv library). Individual points in (e) are jittered along the *x*-axis to increase visibility. See the online article for the color version of this figure.

when the other two components activated (see Figure 12a). 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 each device at least four 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 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 an unusually short $t_{A \to B}$. Before proceeding to the main task, participants had to correctly answer comprehension check questions.

All participants faced each of the five problem types twice, once as detailed in Figure 11 and once with the labels and locations of A and B reversed. Thus, there were 10 within-subjects problems in total. On each problem, participants tested the device 12 times, generating the 12 clips in a random order. For each problem they made three causal judgments. They made their first judgment after the sixth test, their second after all 12 test, and a final judgment after seeing a visual summary of the timelines of the tests they had performed (see Figure 12b). Participants gave their causal judgments by distributing 100 percentage 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.

Results

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 orthogonal contrast codes comparing: (a) (*chain1*, *chain2*) to (*fork1*, *fork2*, and *fork2rev*); (b) *chain1* versus *chain2*; (c) *fork1* to *fork2*; and (d) (*fork1*, *fork2*) to *fork2rev*, matching the normative predictions described above. The four regressions are summarized in Table 3. Judgments for all three responses differed by device type, with the size of these differences increasing for the judgments made after seeing 12 compared with six clips, and after seeing the visual summary relative to before. For instance, after six clips participants assigned 10.2% more percentage points to the chain diagram when the true structure was a chain, increasing to 22.2% after 12 clips 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 with *fork2rev*. However, on the judgment after 12 clips, 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 again equally favored for each of these two devices.

Theoretically, *chain2* and *fork2* are the hardest cases to distinguish because both causal relationships are noisy. Given the substantial difference in judgments between chains and forks overall, we also checked is participants picked up on this, more subtle difference. We found a significant difference in judgments on these two problems at the <.001 level, even after six clips (Table 3, 5th row) demonstrating that participants still distinguished these patterns on average without the aid of visual summary.

Discussion

In sum, we found that participants were able to distinguish between a fork and a chain based on the variability and correlation in event timings alone. However, participants found this inference
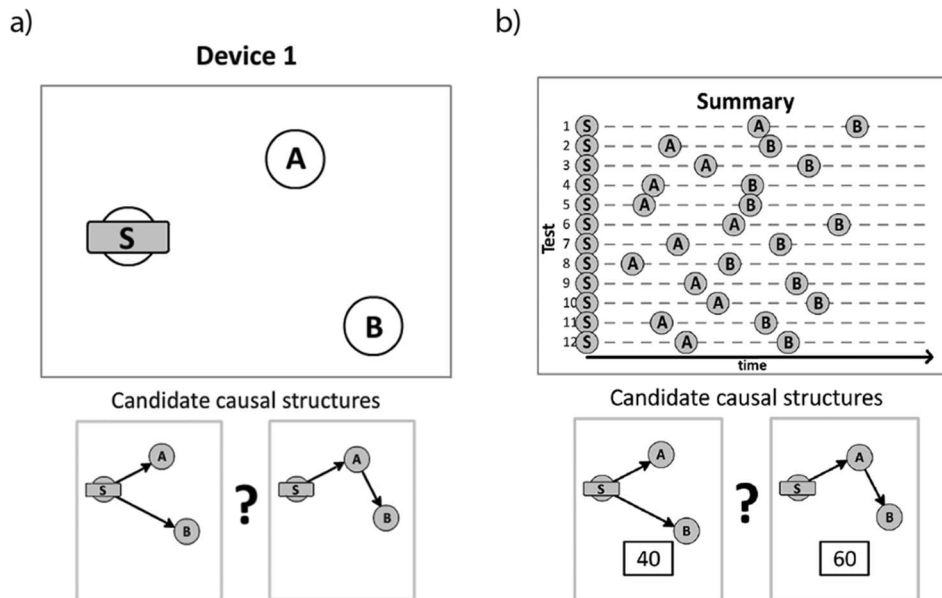

Figure 12. Experiment 3 interface. (a) Testing the device. (b) Viewing a visual summary.

more difficult 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. Furthermore, participants were at chance for *fork1* until viewing the visual summary suggesting that the reliable t_{SB} was hard to detect given the variability in the intervening t_A.

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 11c, 2nd vs. 4th row), t_{SB} is actually also slightly more variable under *fork2* than *chain2* (Figure 11d). Thus, if participants focused on t_{SB} it is not surprising that they 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* relative to *fork2*, is suggestive that participants may have been particularly tuned to monitoring the successive intervals rather than the overall interval (see Figure 13). We examine this idea in more detail in Experiment 4.

Experiment 4: Replication and Extension

From the delay$_I$ model perspective, variability in $t_{S \to A}$ is crucial for distinguishing forks from chains. If t_B "inherits" the variability in t_A such that an early/late t_A is followed by an early/late t_B then this is evidence that the device is a chain. If it does not, this is evidence that the device is a fork. Thus, in the absence of variability in t_A there is no signal to pass on, and the delay$_I$ model cannot strongly favor one model over the other. However, it is also plausible that learners might rely on the variability of t_{AB} or t_{SB} independent of the other timings as cues to structure. For example, one might reject the chain if t_{AB} is "too variable," irrespective of t_{SA}, in which case, systematic chain and fork preferences could still emerge with reliable $t_{S \to A}$. While Experiment 3 focused on cases with variable $t_{S \to A}$, Experiment 4 completes the set of combina-
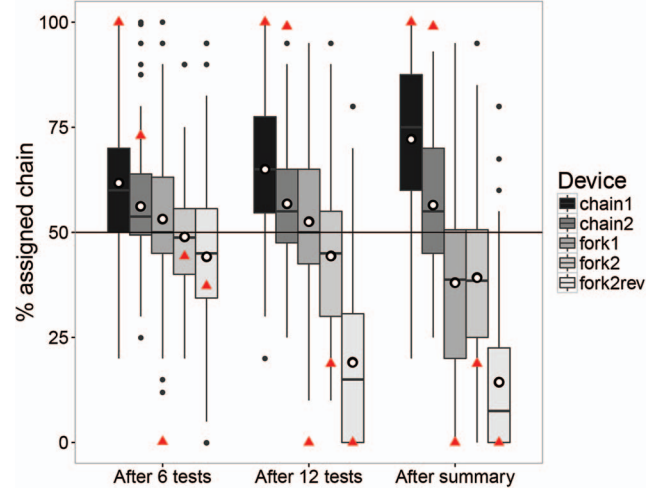

Figure 13. Experiment 3: Judgments for the different devices. Box plots show participants' median and upper and lower quartiles, participants with judgments $\pm >1.5$ interquartile range are plotted separately. White filled black circles represent participant means. Red triangles represent the delay$_I$ model's posteriors. See the online article for the color version of this figure.

tions of variable/reliable connections to better understand which properties of the intervals participants are sensitive to when inferring causal structure.

Method

Participants and materials. Sixty participants (20 female, $M_{age} = 35.5$, $SD_{age} = 9.9$), recruited from Amazon Mechanical Turk, took part in this experiment. The task took 30.6 min ($SD = 15.4$) and participants were paid at a rate of $6 per hour.

Stimuli. We used the same generative stimuli as in Experiment 3 but included four additional devices *chain3*, *chain4*, *fork3*, and *fork4* (see Figure 11, bottom half). *Chain3* and *chain4* have the same $t_{A \to B}$ variability as *chain1* and *chain2,* respectively, but have a reliable instead of a variable $t_{S \to A}$. Likewise, *fork3* and *fork4* have the same $t_{S \to B}$ variability as *fork1* and *fork2,* respectively, but also have a reliable rather than a variable $t_{S \to A}$.

Model predictions. According to the delay$_I$ model, the reliable $t_{S \to A}$ delay makes it hard to detect whether the timing of B is dependent on A. Where variability is also low for $t_{A \to B}$ (*chain3*), the delay$_I$ model does form a preference for the chain structure. However, for *fork3* it requires all 12 data points to form a very weak preference for the fork. It forms no strong preference for *chain4, fork4* where the variability in t_{AB} or t_{SB} overwhelms any dependence. The high reliability of the shorter t_{AB} means the pattern is still more likely on average to have been generated by a chain.

For *chain4* and *fork4*, the variability in t_{AB} renders the correlation, or absence of correlation, between t_{SA} and t_{SB} undetectable from 12 data points.[17] Thus, the delay$_I$ model predicted that participants would be highly uncertain about these additional problems. If, on the other hand, participants relied on the variability

Table 3

Experiment 3: Main Effects and Planned Comparisons for Participants' Responses

	After 6 tests	After 12 tests	After summary
Main effect (LR)	86***	334***	378***
	Planned contrasts		
Intercept	52 ± .9%***	46 ± .9%***	42 ± .9%***
1. Chains vs. Forks	10.2 ± 1.3%***	22.2 ± 1.3%***	33.8 ± 1.6%***
2. Chain1 vs. Chain2	2.7 ± 1.0%*	4.1 ± 1.1%**	7.8 ± 1.3%***
3. Fork1 vs. Fork2	2.1 ± 1.0%	4.0 ± 1.1%**	−.6 ± 1.3%
4. Forks1&2 vs. Fork2rev	4.5 ± 1.1%**	19.6 ± 1.3%***	16 ± 1.5%***
	Additional		
5. Chain2 vs. Fork2	3.6 ± .97%***	6.2 ± 1.2%***	8.6 ± 1.3%***

Note. For main effects we report the likelihood ratio (LR) for a model with device type as predictor relative to a model with just an intercept. For all contrasts we report the size of the effect (%) ± standard error, and level of significance after applying a Bonferroni correction for the total of five contrasts.

* $p < .01$. ** $p < .002$. *** $p < .0002$.

---

[17] Note that in all these cases the true generative model is identifiable in principle, but reliable recovery requires considerably more data.

of the successive intervals, we would expect to see some degree of fork preference if t_{AB} is highly variable. To assess the impact of variability in t_{SA} on structure judgments, we compared judgments between the old and new devices that differed only in terms of $\sigma^2(t_{SA})$: *chain3* against *chain1*, *chain4* against *chain2*, *fork3* against *fork1*, and *fork4* against *fork2*.

Procedure. As in Experiment 3, participants made three judgments about each device. Because we increased the number of devices from five to nine, participants only faced each device once instead of twice, so as to ensure that the experiment was not too long.

Results

Replication of Experiment 3. We found highly similar patterns of judgments for the same devices as presented in Experiment 3. Across the three judgments and five repeated device types, there was a correlation of $r = .98$ between the averaged judgments in Experiment 3 and Experiment 4. We also replicated the same significant effects for the same planned contrasts (as in Table 3). The exceptions are that in Experiment 4, contrasts between *chain1* and *chain2* after 12 tests, *fork1* and *fork2* after 12 tests, and *forks1* and *fork2* and *fork2rev* after summary are no longer significant. As in Experiment 3, participants assigned significantly fewer percentage points to the fork for *fork1* than *fork2* in their two responses before the visual summary contrary to the delay$_I$ model's predictions. Despite very similar patterns, there was a significant global increase in preference for the chain hypothesis across the problems shared between experiments of $M = 4.3\%$, $(SD = 3)$, $t(14) = -5.8$, $p < .001$ (compare white circles with yellow squares in Figure 14).

Extension to new structures. To assess what impact reducing the variability of $t_{S \to A}$ had on participants' judgments, we compare the four *new* devices with their counterparts that only differ in terms of the reliability of $t_{S \to A}$ as described in the Model Predictions section above. We find that participants' chain preference is diminished for *chain3* relative to *chain1* across all three responses (after six tests: $t(59) = 3.2$, $p = .002$; after 12 tests: $t(59) = 2.1$, $p = .04$; and after summary: $t(59) = 3.5$, $p < .001$). Participants' chain preference is increased for *fork3* relative to *fork1* for responses after 12 tests, $t(59) = -2.2$, $p = .02$ and after summary: $t(59) = -4.2$, $p < .001$). *Chain4* and *fork4* responses did not differ significantly from those for *chain2* and *fork2* on any of the three responses. This supports the claim that variability in $t_{S \to A}$ does play an important role in participants' successful inference of the true generative structure.

Overall, participants assigned considerably more than 50% points to the chain hypothesis for *chain3*, *chain4*, *fork3*, and were ambivalent on average about *fork4*. This differs from the predictions of the delay$_I$ model which is neutral about *chain4* (Figure 11e). Another way of unpacking this pattern is that participants made similar judgments for *chain3*, *chain4*, and *fork4*, perhaps simply reflecting a default preference for a chain given staggered activations. Only in the case of *fork4*, with its particularly extreme t_{AB} variability, did they somewhat overturn their chain preference.

Discussion

In this experiment we replicated the findings of Experiment 3 and additionally showed that people do not form a preference for a fork in the absence of significant variability in $t_{S \to A}$, as predicted
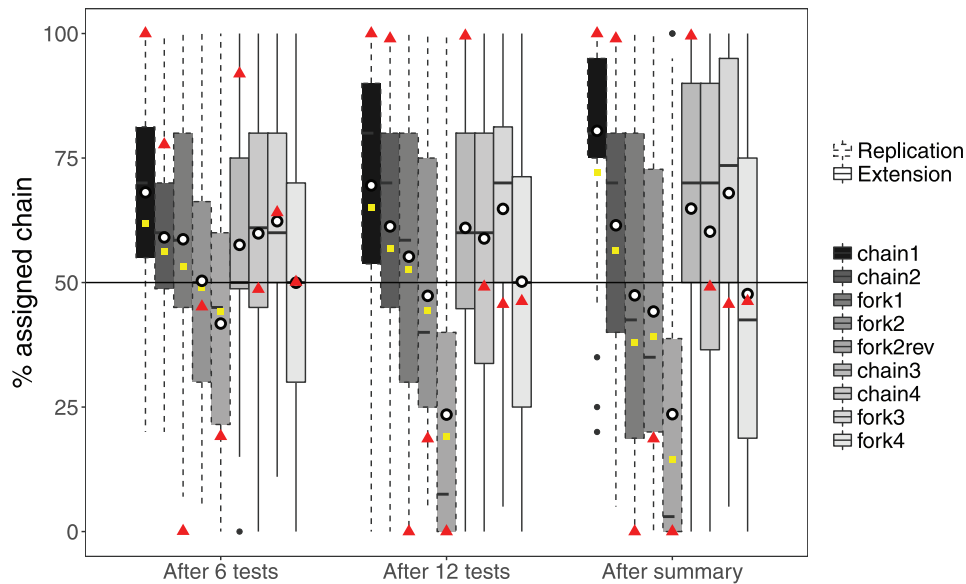
Figure 14. Experiment 4: Judgments for the different devices. Box plots show participants' median and upper and lower quartiles. Dashed borders indicate replication of Experiment 3. Full borders indicate extension to new problems. Yellow squares indicate average judgments from Experiment 3. White filled black circles represent participant means. Red triangles represent the delay$_I$ model's posteriors. See the online article for the color version of this figure.

by the delay$_I$ model. However, there was also a modest overall increase in chain preference compared with Experiment 3. This might have been a consequence of exposure to a different set of devices, with the additional problems' lack of a predictive signal nudging participants away from focusing on this signal. However, it could also be a consequence of population differences. The online population we tested in Experiments 1, 2, and 4 is markedly older and more educationally diverse than the undergraduate psychology students tested in Experiment 3, and may have performed the task under less ideal experimental conditions that may have made it more difficult to perceive the subtler differences in event timing.

Among the new devices unique to Experiment 4, participants were neutral about *fork4* while favoring the chain for the other three. Inspecting Figure 11 we see that the most substantial difference between *fork4* and *chain4* was that the variance of t_{AB} was substantially higher (due to its actually being caused by S). This suggests that the level of variability under the chain hypothesis may still have nudged participants away from the belief that this device is a chain.

We note also that the delay$_P$ model would form a preference for the chain based on *fork3*—because this evidence is particularly consistent with their being a single parameterization for two sequential causal delays. Thus, while we have shown that participants can learn that there are different delays, they may still tend to favor the structure that can explain the results more parsimoniously. In this case a chain with shared delay parameters can capture the evidence more parsimoniously than a model with separate delay parameters.

Overall, our Bayesian delay$_I$ model did a reasonable job capturing participants' structure judgments (see Figure 14). However, there is also evidence that participants used the timing data in a more heuristic way. First, participants' judgments were weaker than what the normative model predicts. Second, the delay$_I$ model strongly favored the fork structure after seeing only a few clips from *fork1*, while participants remained at chance for this problem until having seen the visual summary at the end.

General Discussion

In this article we propose a normative account of how the *when* informs the *why*: how temporal delays between events allow inferences about their underlying causal structure. We used this account as a framework to closely explore human patterns of causal structure inference from time. We distinguished inference from event order from sensitivity to timing information. We focused on abstract structure induction to minimize the influence of domain-specific prior beliefs and to isolate, as far as possible, domain general causal learning patterns and preferences.

In Experiment 1, we found that participants' one-shot structure judgments were shaped by event order and sensitive to how many different order patterns each candidate structure could generate. Participants were also influenced by the exact event timings, preferring structures that could have produced the evidence with more equal cause–effect delays.

In Experiment 2, we explored how people integrate information from multiple observations of a device. We elicited prior, likelihood, and posterior judgments, allowing us to explore participants' deviations from normative inference under our initial assumptions.

We found that some participants' priors were compatible with uniformity at the level of types of structure rather than the options we presented them with. Furthermore, while we initially assumed an equal likelihood for all evidence patterns a device was capable of producing, participants were not indifferent in this way. Participants assigned more likelihood to patterns implying similar cause–effect delays occurred across connections, in line with a general preference for causal regularity. At the individual level, participants differed in how sensitive they were to the exact delays in the evidence they experienced. With these factors in place, we were able to capture participants' posterior judgments to a high degree of quantitative accuracy.

Experiments 3 and 4 focused on the role of event timing in distinguishing between structures that are indistinguishable by temporal order alone. 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.

In sum, our modeling and behavioral findings present the following picture of human structure inference from temporal information: (a) when both delay expectations and evidence are very limited, people rely on order information; (b) as evidence accrues, they use reversals to rule out structures as captured by the order model and favor structures that imply similar delays as captured by the delay$_P$ model; and (c) if there is sufficient evidence, they are also able to use a subtler causal signal, the variation and correlation in event times as captured by the delay$_I$ model. Overall, we found that people are capable of using temporal information at a level of sophistication that depends on necessity and the available data.

Alternative Timing Models

Ideal probabilistic structure inference involves maintaining a probability distribution over all candidate hypotheses. Although Experiments 3 and 4 only featured two hypotheses, considering all hypotheses is infeasible in the general case as there is a theoretically infinite number of possible models. There have been several recent proposals suggesting that people maintain a single candidate causal model at a time, and switch to another candidate when their current model proves strongly incompatible with the observed evidence (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Bramley et al., 2017). 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 used simpler statistics to identify the generative model without computing the predictions under both structures at once. We now propose several specific aspects of the clips that participants might have been sensitive to.

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 extent that the timing of A predicts the timing of B. In the current context this comes down to a correlation between t_{SA} and t_{SB}, hereafter cor(t_{SA}, t_{SB}). If 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 11a). 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 places high 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 clip, storing them, and comparing their relationship across multiple clips. It is well-established that there are strong limitations on explicit attention and short-term memory (STM) 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. 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_{XY}^k and the one before t_{XY}^{k-1}, 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—$cor(t_{SA}, t_{SB})$, $\sigma^2(t_{SA})$, $\sigma^2(t_{SB})$, $\sigma^2(t_{AB})$, $APD(t_{SA})$, $APD(t_{SB})$, and $APD(t_{AB})$—assigns a value to the evidence seen at each time point by each participant in Experiments 3 and 4. 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 for all devices (except *fork2rev*) 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 structure.

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 from Experiments 3 and 4 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 1 *SD* increase in the value of each independent variable. We entered the predictors detailed in Table 4.

Figure 15 depicts the models selected by the stepwise procedure for the first, second, and third judgments, respectively. Table 5 shows the bivariate correlations between all the predictors we considered.

For the first judgment, occurring after six tests, one of the eight predictors was chosen and the rest were eliminated. 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 observation to the previous one (i.e., $APD(t_{AB})$). The fact that the intercept was »0 indicates a baseline preference for the chain that could be overturned by high $APD(t_{AB})$.

For the second judgment which occurred after having seen all the evidence, the actual variance of t_{AB} was selected rather than its sequential proxy. This time the sequential measure of variance in t_{SB} model additionally contributed to the prediction.

We see a similar pattern for the final judgment participants made after having seen the visual summary. However, now $APD(t_{AB})$, $\sigma^2(t_{AB})$) and $APD(t_{SB})$) all contributed more substantially to the fit.

On all judgments, but particularly those before the visual summary, the strongest single predictor was either $\sigma^2(t_{AB})$, or $APD(t_{AB})$ with apparent variance contributing to assignment of more points to the chain (see Table 5). Corroborating this, Figure 16 visualizes participants' pattern of judgments by device and compares this against $-\sigma^2(t_{AB})$ and delay$_I$. On the first two judgments in particular, only $-\sigma^2(t_{AB})$ captures participants' ambiva-

Table 4

Experiments 3 and 4: Alternative Predictors

| Name	Description
Intercept	Positive value captures overall preference for chain, negative for fork.
$cor(t_{SA}, t_{SB})$	The correlation between the delays $t_{SA}^{1:k}$ and $t_{SB}^{1:k}$.
$\sigma^2(t_{xy})$	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 $\sigma^2(\widehat{t_{AB}})$).
$APD(t_{xy})$	Average pairwise difference. A sequentially computed proxy for variance: The difference in activation time on current test compared with previous test for example t_{xy}^k and t_{xy}^{k-1} summed up over tests 1:K. e.g. for t_{AB} after six trials this is $APD(t_{AB}) = \Sigma_{k=2:6} t_{AB}^k - t_{AB}^{k-1}$. As with the variance, we entered the APD for each interevent interval.
Delay$_I$	The posterior probability of the chain according to delay$_I$.
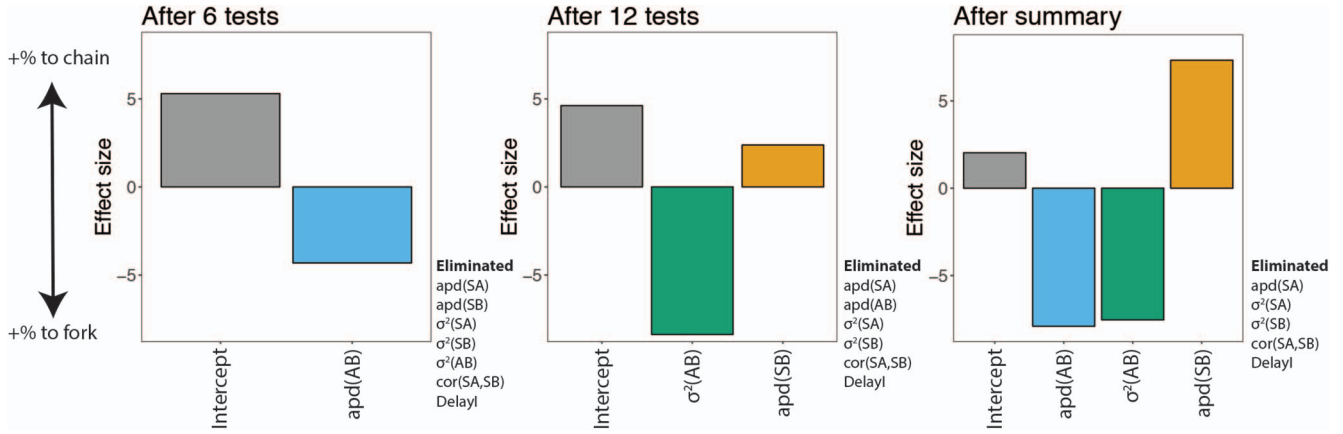
Figure 15. The models resulting from stepwise model selection to all 1,040 first, second, and final judgments in Experiment 3. Plots show the selected predictors' fixed effects (i.e., the β values) with consistent colored fills across subplots to aid comprehension. All predictors were z scored, and the dependent variable was centered (so that a prediction of 0 corresponded to assigning 50% to the chain and 50% to the fork). Effect sizes are interpretable as differences in percentage assigned to the chain moving one standard deviation up on the independent variable. See the online article for the color version of this figure.

lence between *chain2* and *fork1* and confidence about *fork2*. In no case was the Bayesian posterior was selected as part of the final model.

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 were more strongly influenced by

both the putative chain delay t_{AB} and putative fork delay t_{SB}, consistent with the idea that where t_B does covary with t_A this pops out clearly when viewing the summary timelines side by side (Figure 12b). Only after seeing this visual summary did participants "get" *fork1* while the $\sigma^2(t_{AB})$ still predicts that participants should be undecided in this case.

Conjunctive influence. In Experiments 1 and 2, we instructed participants that the collider structure was conjunctive and 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, some participants

Table 5

Experiments 3 and 4: Bivariate Correlations Between Judgments, Delay$_I$, and Heuristic Measures

	Delay$_I$	cor(t_{SA}, t_{SB})	$\sigma^2(t_{SA})$	$\sigma^2(t_{SB})$	$\sigma^2(t_{AB})$	APD(t_{SA})	APD(t_{SB})	APD($t_{\widehat{AB}}$)
			First judgment (after 6 tests)					
% Chain 1st	.13	.12	.01	−.06	−.19	.00	−.04	−.19
Delay$_I$		.59	−.08	.28	−.32	−.08	.35	−.45
cor(t_{SA}, t_{SB})			.14	.11	−.59	.16	.22	−.58
$\sigma^2(t_{SA})$				.21	.18	.83	.25	.23
$\sigma^2(t_{SB})$					.58	.19	.87	.47
$\sigma^2(t_{AB})$						.15	.44	.86
APD(t_{SA})							.34	.30
APD(t_{SB})								.43
			Second and third judgments (after all 12 tests)					
% Chain 2nd	.22	.22	−.05	−.11	−.29	−.03	−.06	−.27
% Chain 3rd	.40	.38	−.09	−.04	−.37	−.07	.03	−.39
Delay$_I$		.92	−.05	.26	−.59	−.03	.32	−.61
cor(t_{SA}, t_{SB})			.21	.20	−.49	.16	.13	−.47
$\sigma^2(t_{SA})$				.25	.20	.75	.10	.07
$\sigma^2(t_{SB})$					.55	.23	.75	.41
$\sigma^2(t_{AB})$						.13	.33	.72
APD(t_{SA})							.40	.33
APD(t_{SB})								.57

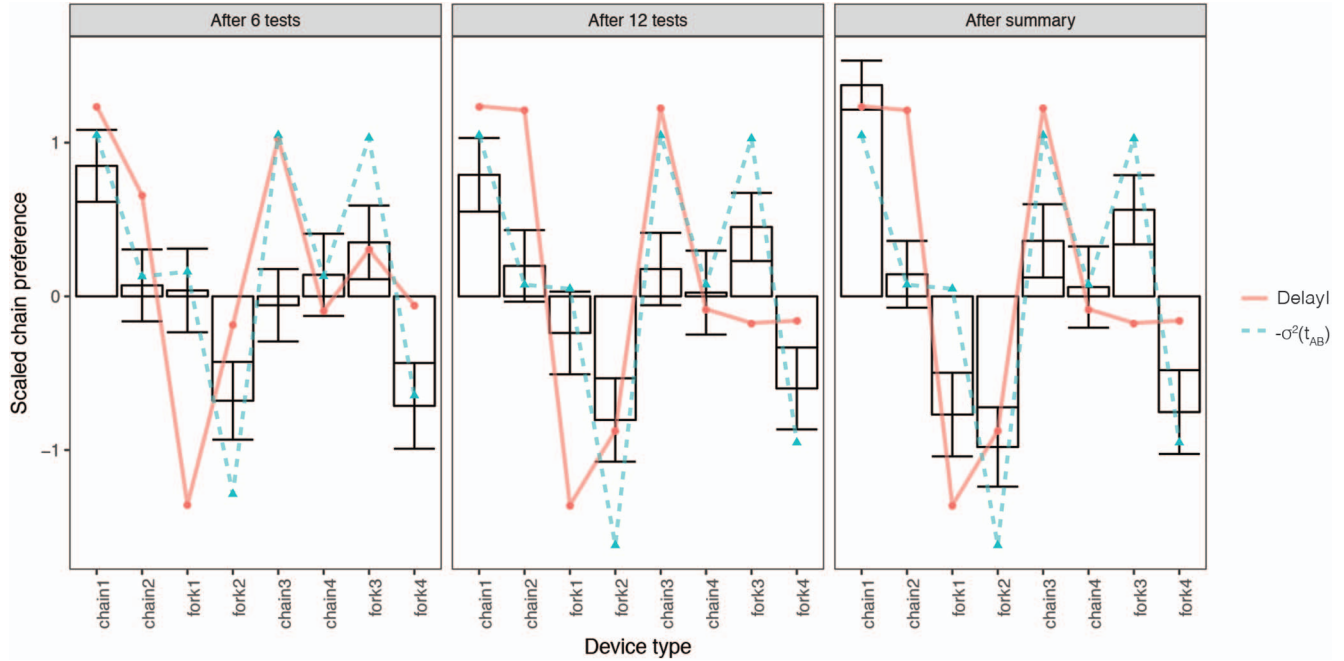*Note.* APD = Average pairwise difference.

Figure 16. Comparison of participants patterns in Experiment 4 with delay$_I$ predictions and $-\sigma^2(t_{AB})$. Judgments and predictors are rescaled to have a mean of 0 and standard deviation of 1 across the three judgments for comparability. See the online article for the color version of this figure.

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 posterior probabilities to the collider even after observing clips where one of its' cause components occurred after the effect, as well as nonzero likelihoods, under the collider, for these patterns. 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 task 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 later arriving causal influence was what determined when the effect would occur. This could be understood as the earlier-arriving causal influence waiting in a "buffer" for the later 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. This would explain participants strong preference for a collider when *A* and *B* co-occurred in Experiment 1, Device a. Additionally, people might find conjunctive influence 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.

Causal and Temporal Perception

We looked only at a narrow range of time intervals in the current studies, with trials never lasting more than around three seconds. For these short intervals, existing causal beliefs have been shown to shape and 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, 2013). This suggests that at this temporal grain, experience is still somewhat under construction (Dennett, 1988), scaffolded by preexisting expectations about causal structure.

Having formed an impression that a device has a certain structure, a subsequent observation may be perceived as consistent with that structure even though it was actually inconsistent. This might occur more often if the distortion required to make it consistent is very small. Such potential distortion 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. A first step toward this might be to simply add perceptual noise in proportion to the temporal delays. 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 structural hypotheses, 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 on the order of minutes (e.g., the effects of ingredients and temperature changes in cooking) or weeks (e.g., policy changes ef-

fects on social issues) are likely a substantially different cognitive processes. There is more time to explicitly reason about and compare hypotheses but also more intervening demands on attention that may limit fine grained timing.

The Blessing of Variability

Our experimental design highlights an interesting and counterintuitive property of temporal causal structure 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-based inference. A similar principle applies in the absence of revealing order information (cf. Young & Nguyen, 2009). It is actually the variability in delays that allows for the correlation signal that the delay$_t$ model uses 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 without some prior expectation of the lengths of the different delays.

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 somewhat unreliable (Bramley et al., 2015; Fernbach & Sloman, 2009).

One Model at a Time?

Some participants in Experiments 1 and 2 assigned more probability to structures that, if they produced the evidence, did so with cause–effect delays that were of roughly equal length. For instance, participants assigned more probability to the fork than the collider based on Clip 2, and more to the collider than the fork based on Clip 6 in Experiment 1. Additionally, participants' likelihood assignments over qualitative patterns in Experiment 2 was clearly consistent with a preference for equal delays. However, judgments in Experiments 3 and 4 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., 2017). Because seeing several events that always occur in the same order *ceteris paribus* is most naturally perceived as a chain, participants may have begun the problems in Experiments 3 and 4 with a sense of watching a causal chain, which could be gradually overturned in the cases where there was another more plausible hypothesis available (of the device as a fork).

Building Richer Causal Representations

While CBNs provide a useful framework for building theories about causal cognition, they are not rich enough to explain central aspects of causal cognition such as detailed mechanism knowledge or mental simulation (Gerstenberg et al., 2015; Mayrhofer & Waldmann, 2015; Sloman & Lagnado, 2015; Waldmann & Mayrhofer, 2016). People's causal representations are likely to lie somewhere in between a compact statistical map (a CBN) and a scale model of the physical world (Gerstenberg & Tenenbaum, 2017; Goodman, Tenenbaum, & Gerstenberg, 2015; Lake, Ullman, Tenenbaum, & Gershman, in press; Tenenbaum et al., 2011; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). 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 delay model represents a step toward capturing the ways in which human causal cognition goes beyond contingencies.

Conclusions

We have shown in four experiments that people form systematic beliefs about causal structure based on temporal information. We can capture 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. 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 initial belief as they observe more evidence about the system by considering what structural changes from this default better explain what happened when. These results contribute to our scientific 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 how we learn and represent the world.

References

Addyman, C., French, R. M., & Thomas, E. (2016). Computational models of interval timing. *Current Opinion in Behavioral Sciences, 8,* 140–146.

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,* 299–352.

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

Bechlivanidis, C. (2015). *The arrow of time through the causal lens: When causal beliefs determine temporal order* (Unpublished Doctoral dissertation). University College London, London, England.

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

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

Blaisdell, A. P., & Waldmann, M. R. (2012). Rational rats: Causal inference and representation. In E. A. Wasserman & T. R. Zentall (Eds.), *Oxford handbook of comparative cognition.* http://dx.doi.org/10.1093/oxfordhb/9780195392661.013.0011

Bonawitz, E. B., 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., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review, 123,* 301–338.

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., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 236–242). Austin, TX: Cognitive Science Society.

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

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

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

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

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

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning, 12,* 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,* 3–5.

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

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. New York, NY: Oxford University Press.

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

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

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* 62.

Clayton, N., & Dickinson, A. (2006). Rational rats. *Nature Neuroscience, 9,* 472–474.

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

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations. *The Journal of general psychology, 125,* 245–261.

Courville, A. C. (2006). *A latent cause theory of classical conditioning* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Courville, A. C., Daw, N., Gordon, G. J., & Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In S. Thrun, L. Saul, & B. Schoelkopf (Eds.), *Advances in neural information processing systems* (pp. 977–984). Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2002). Similarity and discrimination in classical conditioning: A latent variable account. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 313–320). Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294–300.

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,* e57410.

de Finetti, B. (1975). *Theory of probability: A critical introductory treatment* (Vol. 6). New York, NY: Wiley.

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. In *Proceedings of the 34th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.

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

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 678.

Fraley, C., Raftery, A. E., Murphy, T. B., Scrucca, L. (2012) *Normal mixture modeling for model-Based custering, classification, and density estimation (mclust Version 4 for R),* Technical Report No. 597, Department of Statistics, University of Washington, Seattle, Washington.

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.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107,* 289.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis. Texts in statistical science series*.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior, 40,* 255–268.

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

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldman (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). New York, NY: Oxford University Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32,* 108–154.

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). Cambridge, MA: MIT Press.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37,* 620.

Greville, W. J., & Buehner, M. J. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition, 35,* 444–453.

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

Greville, W. J., & Buehner, M. J. (2016). Temporal predictability enhances judgements of causality in elemental causal induction from both observation and intervention. *The Quarterly Journal of Experimental Psychology, 69,* 678–697.

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 & Cognition, 41,* 904–916.

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

Griffiths, T. L. (2005). *Causes, coincidences, and theories* (Unpublished doctoral dissertation). Stanford University, Stanford, CA.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 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,* 382–385.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 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,* 803–831.

Hartigan, J. A. (2012). *Bayes theory.* Berlin, Germany: 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, 48,* 400–407.

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

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,* 1185–1243.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80.

Lagnado, D. A., & Sloman, S. A. (2002). Learning causal structure. In W. D. Gray and C. Schunn (Eds.), *Proceedings of the 24th annual meeting of the cognitive science society* (pp. 560–565). Hillsdale, NJ: Erlbaum.

Lagnado, D. A., & Sloman, S. A. (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,* 451–460.

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

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (in press). Building machines that learn and think like people. *Behavioral and Brain Sciences.*

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

Lee, P. M. (2012). *Bayesian statistics: An introduction.* New York, NY: Wiley.

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

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

MacKay, D. J. C. (1991). *Bayesian methods for adaptive models* (Unpublished doctoral dissertation). California Institute of Technology, Pasadena, CA.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms.* New York, NY: Cambridge University Press.

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

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23.

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

Mayrhofer, R., & Waldmann, M. R. (2016). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science, 40,* 2137–2150.

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,* 277.

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

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). Austin, TX: Cognitive Science Society.

Pacer, M. D., & Griffiths, T. L. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67,* 186–216.

Pavlov, I. P. (1928). Lectures on conditioned reflexes. New York, NY: W H Gantt International Publishers.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann.

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

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14,* 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,* 383.

Rehder, R. (2011). Reasoning with conjunctive causes. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1406–1412). Austin, TX: Cognitive Science Society.

Rehder, R. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72,* 54–107.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe flash and html5/Javascript web experiments. *Behavior Research Methods, 47,* 309–327.

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140,* 109–139.

Rottman, B. M., & Hastie, R. (2016). 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,* 93–125.

Savage, L. J. (1972). *The foundations of statistics.* North Chelmsford, MA: Courier Corporation.

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, 21,* 229–261.

Skinner, B. F. (1938). *The behaviour of organisms: An experimental analysis*. New York, NY: D. Appleton-Century Company Incorporated.

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

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

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

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press.

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,* 1279–1285.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21,* 649–665.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. *Psychology of Learning & Motivation, 65,* 85–127.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20,* 158.

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

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

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

Young, M. E., & Nguyen, N. (2009). The problem of delayed causation in a video game: Constant, varied, and filled delays. *Learning and Motivation, 40,* 298–312.

Appendix

Additional Modeling Details

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 (a) 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 (b) 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

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

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

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

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)

$$p(t_{AE}, t_{BE} | \alpha, \mu) = p(t_{AE}, t_{BE} | \alpha, \mu; t_{AE} = t_{A \to E}, t_{BE} \geq t_{B \to E})$$
$$+ p(t_{AE}, t_{BE} | \alpha, \mu; t_{AE} \geq t_{A \to E}, t_{BE} = t_{B \to E}) \quad \text{(A-3)}$$
$$= p(t_{AE} | \alpha, \mu) \cdot p(t_{BE} \geq t_{B \to E} | \alpha, \mu)$$
$$+ p(t_{BE} | \alpha, \mu) \cdot p(t_{AE} \geq t_{A \to E} | \alpha, \mu) \qquad \text{(A-4)}$$

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

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

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

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

(Appendix continues)

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: (a) 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 (b) the influence of B arrives first overshadowing the influence of A (i.e., $t_{BE} = t_{B \to E}$ but $t_{AE} \geq t_{A \to E}$). Thus, the joint likelihood of a disjunctive (pooled delay) collider can be written as

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

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

Timing Based Likelihoods

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 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 $tt_X \in d_i$ onto the structure of the model $s \in S$

$$p(\mathbf{d} | \lambda, \alpha, \mu; s) = \prod_{i \in 1:n} p(d_i | \lambda, \alpha, \mu; s)$$
$$= \prod_{i \in 1:n} \prod_{t_X^i \in d_i} p(t_X^i - t_{pa(X)}^i | \lambda, \alpha, \mu; s) \tag{A-10}$$

with $p(t_X^i - t_{pa(X)}^i | \lambda, \alpha, \mu; s)$ 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} according to the derivation in above.

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[18]

$$p(\mathbf{d} | s) = \int p(\mathbf{d}, \lambda, \alpha, \mu | s) \, d\lambda \, d\alpha \, d\mu \tag{A-11}$$

$$= \int p(\mathbf{d} | \lambda, \alpha, \mu; s) \cdot p(\lambda, \alpha, \mu | s) \, d\lambda \, d\alpha \, d\mu \tag{A-12}$$

$$= \int p(\mathbf{d} | \lambda, \alpha, \mu; s) \cdot p(\lambda | s) \cdot p(\alpha | s) \cdot p(\mu | s) \, d\lambda \, d\alpha \, d\mu \tag{A-13}$$

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

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 A13) at the sampled points in parameter space

$$p(\mathbf{d} | s) = \int p(\mathbf{d} | \lambda, \alpha, \mu; s) \cdot p(\lambda | s) \cdot p(\alpha | s) \cdot p(\mu | s) \, d\lambda \, d\alpha \, d\mu \tag{A-14}$$

$$\approx \frac{1}{B} \sum_{b=1}^{B} p(\mathbf{d} | \lambda^{(b)}, \alpha^{(b)}, \mu^{(b)}; s) \tag{A-15}$$

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

Markov Chain Monte Carlo Estimation: Experiments 3 and 4

In Experiments 3 and 4, 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"

$$p(\alpha, \theta | \mathbf{d}; s) \propto \frac{p^{\alpha-1} e^{-\frac{q}{\theta}}}{\Gamma(\alpha)^n \theta^{\alpha n}} \tag{A-16}$$

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:

---

[18] Concretely, we used an exponential (0.1) prior for α, an exponential (0.0001) prior on μ, and an exponential (10,000) prior on λ, corresponding to a weak expectation for positive dependence, shorter delays, and frequently occurring independent causes.

[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.
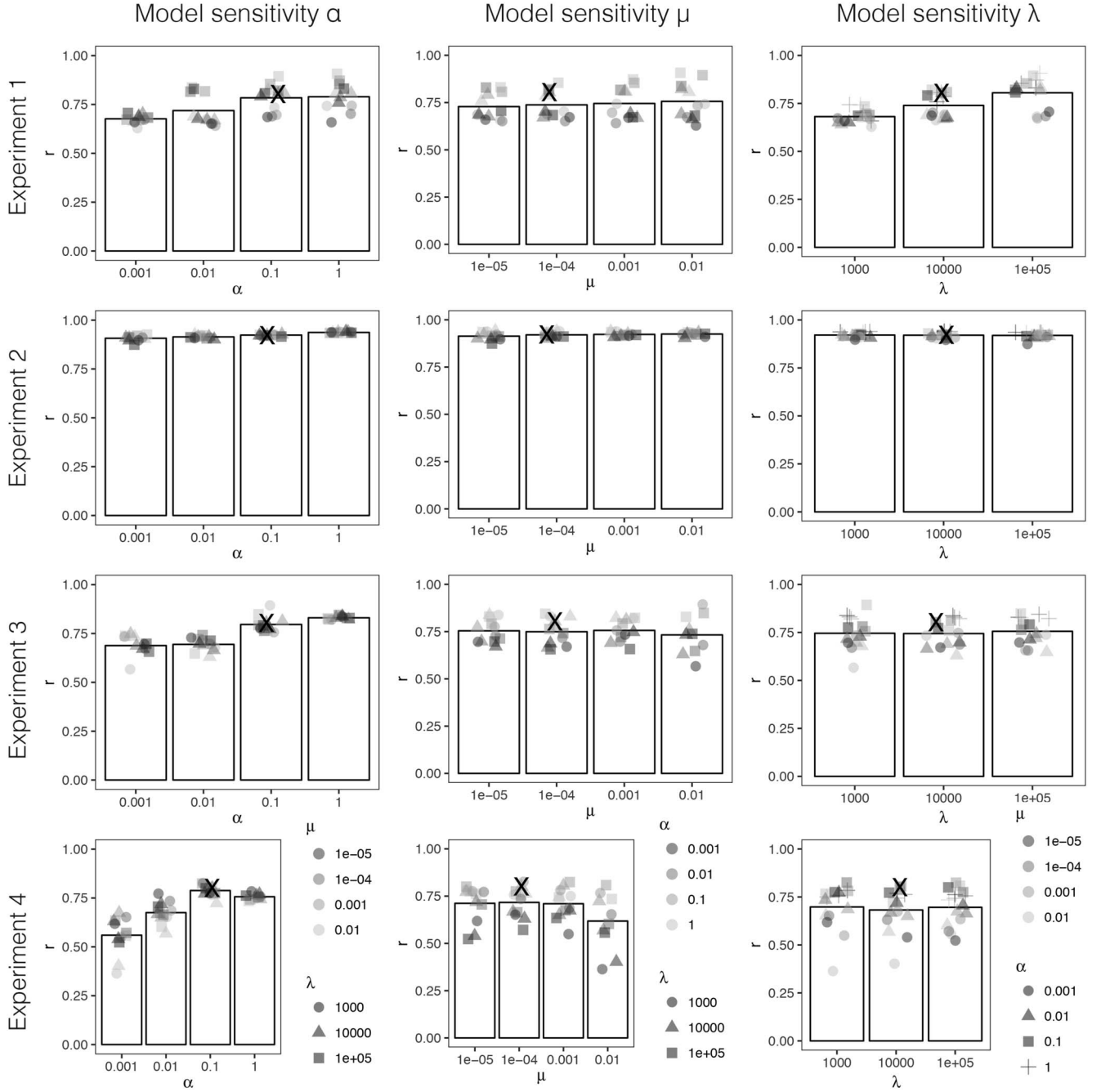
Figure A1.   Sensitivity of model predictions to priors on α, μ and λ. The settings used in the article are denoted by an "X" in each case. Legends apply to all plots in the column.

(*Appendix continues*)
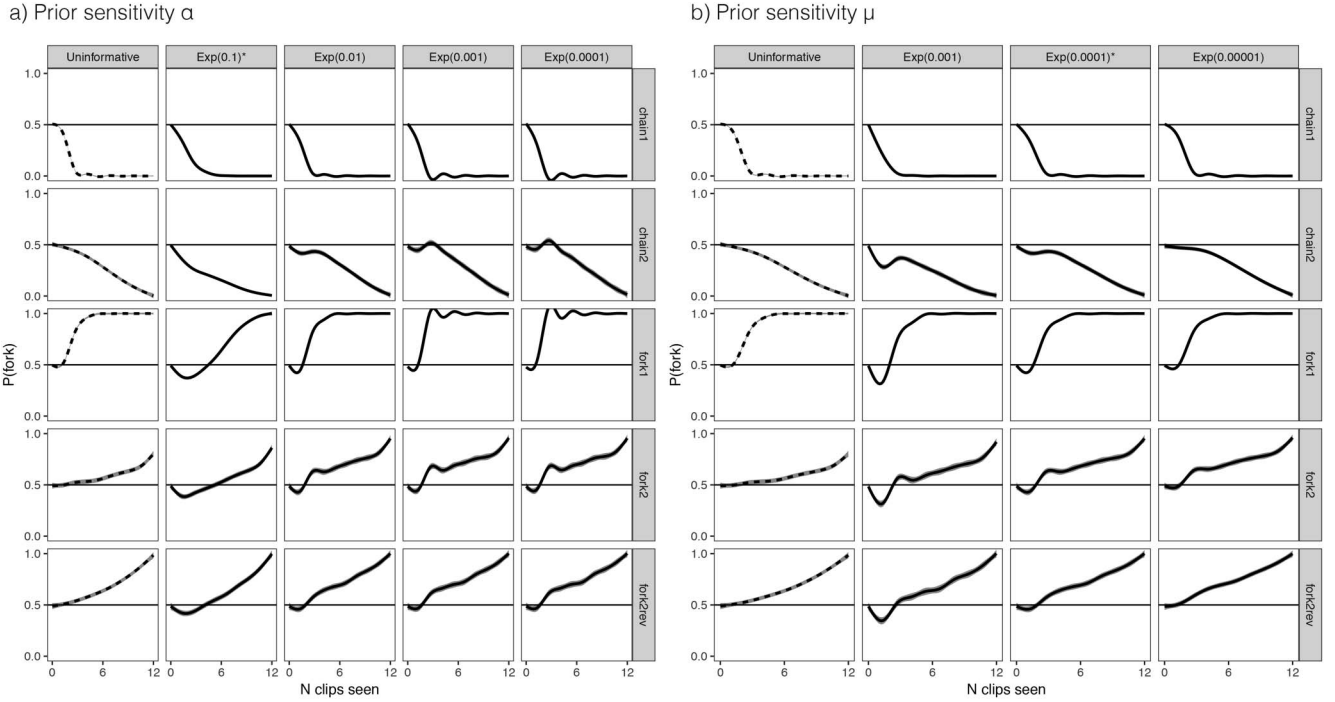
a) Prior sensitivity α

b) Prior sensitivity μ

Figure A2. (a) Sensitivity of α prior on model predictions in Experiments 3 and 4. Left hand column (dashed line) shows predictions using an "improper" uninformative prior. Other columns show predictions under different priors on α. Asterisk indicates the values used for predictions in Experiments 1 and 2. Lines smoothed using the general linear additive model with integrated smoothness estimation gam from R's mgcv library with gray surrounds indicating a 99% confidence interval. The prior on μ for these simulations was exponential (0.0001). (b) Sensitivity of μ prior on model predictions in Experiments 3 and 4. Columns, asterisks, and lines as in a. The prior on α for these simulations was exponential (0.1).

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 ($SD = 50$) and θ ($SD = 10$) 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 first assess the sensitivity of our model fits to our choice of priors on α, μ, and λ before comparing them against the predic-

tions of the Markov Chain Monte Carlo procedure we used to estimate posteriors in Experiments 3 and 4. 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.

Figure A1 shows that we get similar delay model fits for a range of α, μ, and λ values spanning three to four orders of magnitude in each case. In general there is some variability, particularly in Experiment 4 where for half of the problems there is essentially no causal signal, and meaning even weak prior expectations can dominate. Figure A2a and b compare our Simple Monte Carlo estimations directly against our prior-free MCMC procedure. We see 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 in Experiments 1 and 2.