THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

# Continuous time causal structure induction with prevention and generation

**Citation for published version:**
Gong, T & Bramley, NR 2023, 'Continuous time causal structure induction with prevention and generation', *Cognition*, vol. 240, 105530. https://doi.org/10.1016/j.cognition.2023.105530

**Digital Object Identifier (DOI):**
10.1016/j.cognition.2023.105530

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Publisher's PDF, also known as Version of record

**Published In:**
Cognition

**General rights**
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. Oct. 2023

Cognition 240 (2023) 105530

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

# Continuous time causal structure induction with prevention and generation ☆

Tianwei Gong *, Neil R. Bramley

*Department of Psychology, University of Edinburgh, United Kingdom*

ARTICLE INFO

Dataset link: https://osf.io/q8n72/, https://github.com/tianweigong/causal_diamond

*Keywords:*
Causal learning
Time
Prevention
Structure induction
Summary statistics

ABSTRACT

Most research into causal learning has focused on atemporal contingency data settings while fewer studies have examined learning and reasoning about systems exhibiting events that unfold in continuous time. Of these, none have yet explored learning about preventative causal influences. How do people use temporal information to infer which components of a causal system are generating or preventing activity of other components? In what ways do generative and preventative causes interact in shaping the behavior of causal mechanisms and their learnability? We explore human causal structure learning within a space of hypotheses that combine generative and preventative causal relationships. Participants observe the behavior of causal devices as they are perturbed by fixed interventions and subject to either regular or irregular spontaneous activations. We find that participants are capable learners in this setting, successfully identifying the large majority of generative, preventative and non-causal relationships but making certain attribution errors. We lay out a computational-level framework for normative inference in this setting and propose a family of more cognitively plausible algorithmic approximations. We find that participants' judgment patterns can be both qualitatively and quantitatively captured by a model that approximates normative inference via a simulation and summary statistics scheme based on structurally local computation using temporally local evidence.

We naturally think about the world in terms of a progression of events that cause and affect one another. When successful, causal reasoning helps us abstract from our real-time experience to recognize stable causal mechanisms that we can use to explain, predict and sometimes control our environment (Sloman, 2005). However, inferring causal structure in real environments is notoriously challenging, involving a complex interplay between incoming evidence, action, and intuitive theories of how causal influences manifest and link elements of experience like events, objects and variables (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009; Lagnado, 2011).

Two of the basic and well-studied notions of causality are generative and preventative relationships. In a generative relationship, we think of the occurrence of one event as bringing about the occurrence of another. A *generative* causal claim implies the counterfactual that, had the cause not occurred, the effect would not have occurred either. In probabilistic accounts of causal reasoning, generative causality is typically linked with an expectation of positive contingency: The presence of a generative causal variable is associated with an increase in the probability of its effect(s) being present compared to cases where the cause is absent or inactive. The reverse of this is the notion of a *preventative* causal relationship, where we think the occurrence of a causal event as blocking another event from occurring. A preventative causal claim implies the counterfactual that, had the cause not occurred, the effect would have occurred. Probabilistically, we thus expect the presence of a preventative cause to decrease the probability of its effect(s) occurring, compared with cases where the cause is absent or inactive (Cheng, 1997; Griffiths & Tenenbaum, 2005; Sloman, 2005).

The majority of causal learning research has focused on inferences from atemporal evidence, which can be represented in tables of co-occurrence or contingency that reflect the statistical dependencies among a set of variables (Buehner, Cheng, & Clifford, 2003; Cheng, 1997; Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2004; Rottman & Hastie, 2014). This kind of data is central in scientific experimentation, in that it depends on the collection of multiple independent samples (Pearl, 2000; Pearl & Mackenzie, 2018; Zimmerman, 2007). However, an intriguing question regarding human cognition is about how people learn causal relationships from temporal data, given that

☆ **Author Note**

A preliminary analysis of a pilot experiment and Experiment 1 was presented at the 42th Annual Meeting of the Cognitive Science Society (Gong and Bramley, 2020) and the Causal Inference & Machine Learning workshop at the 35th Neural Information Processing Systems conference. We thank Simon Stephan and an anonymous reviewer for many helpful comments. TG is supported by a University of Edinburgh PPLS Scholarship. NB is partly supported by a EPSRC New Investigator Grant (EP/T033967/1).

* Corresponding author.
*E-mail address:* tia.gong@ed.ac.uk (T. Gong).

https://doi.org/10.1016/j.cognition.2023.105530
Received 26 October 2022; Received in revised form 15 June 2023; Accepted 16 June 2023
Available online 16 August 2023
0010-0277/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

we experience the world as one continuous timeline, and that real world causal mechanisms often take time to produce their effects. The temporal setting also allows that multiple events of the same type may occur multiple times to a single individual. This more closely resembles repeated-measure data from a single individual than reasoning from large independent samples. In this setting, people might rely more on "soft" cues (e.g. time, prior knowledge) than the contingency principle (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Understanding how people learn from temporal data is crucial because it not only improves our understanding of the basic mechanisms of human learning, but also clarifies the differences between scientific practices and intuitive causal inference.

Besides this, studies of atemporal causal learning (Buehner et al., 2003; Cheng, 1997; Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2004; Rottman & Hastie, 2014) as well as recent studies of temporal causal learning (Bramley, Gerstenberg, Mayrhofer and Lagnado, 2018; Buehner & McGregor, 2006) typically focus on one type of causal relationship at a time. In contrast, this paper aims to investigate how one can learn preventative and generative relationships where both are in play at once. Can people identify what is causing and what is preventing an effect *despite*, and perhaps even *because* of the ways that such causal influences intertwine and interact in time. Although this may sound like a "niche" scenario, it is actually very common. To illustrate such an everyday situation: Suppose you adopt a cat that, while adorable, frequently urinates outside its litter. You would like to understand why and learn to prevent this behavior before she completely ruins your soft furnishings.[1] Identifying the causes of the problem peeing, not to mention an effective pee-prevention strategy is nontrivial and might require considerable thought and experimentation. Perhaps you notice the cat rarely pees inappropriately when playing with its teaser. However it is unclear if teaser is an effective preventer, because the times of day she is encouraged to play with it may be different from those when she pees. Intuitively, diagnosis becomes easier if you can exploit the moments when you know she tends to urinate to test whether the teaser is an effective preventer. For instance, if she often urinates around 7 a.m, you could try introducing her teaser around this time. Alternatively, you might consider encouraging her to drink water to stimulate additional need to urinate a little before the time she more habitually plays with her teaser. In this way you might leverage either an established baseline expectation or an established generative cause (extra water) to facilitate your preventative investigation.

The example above shows, firstly, that temporal expectations are necessary to make sensible causal inferences (Bramley, Gerstenberg, Mayrhofer et al., 2018; Buehner & McGregor, 2006; Greville & Buehner, 2010; Lagnado & Sloman, 2006). In this case we need some sense of when the cat usually pees inappropriately, as well as an expectation of how long it takes for water to pass through its body. Secondly, it is likely that generative and preventative influences *interact* in terms of how they reveal or obscure one another (Lombrozo, 2010; Rottman, 2016). The existence of either a regular base rate occurrence of an effect, or of effects generated by a known generative cause with regular delays, makes it possible to form a strong expectation against which we can test preventative causes.

In this paper we distill these reasoning patterns into a task and a rational analysis that aim to examine: (1) whether people can use temporal knowledge to learn causal systems that include both generative and preventative causes, (2) how the regularity or predictability of the base rate occurrence of an effect of inference affects the learning process, and (3) whether there are interactions between learning different types of causes.

Apart from establishing what factors influence temporal causal learning, we also want to know *how* people learn, i.e. what kind of inference process can capture human judgments. Causal Bayesian Networks (CBNs) are an established mathematical framework representing and reasoning about causal structure giving rise to observations (Allan, 1980; Pearl, 2000; Rottman & Hastie, 2014). In the psychology of causal reasoning, they have served as a computational-level norm (Marr, 1982) allowing researchers to investigate how the cognitive processes of causal induction approximate or deviate from ideally reverse engineering the generative causal mechanism most likely to be responsible for one's observations. Accordingly, a number of process-level models have been proposed (Bramley, Dayan, Griffiths and Lagnado, 2017; Davis & Rehder, 2020) that each capture some of the ways human performance departs from this kind of Bayesian ideal. However, CBNs and extant process-level models do not describe the role of continuous-time information in human causal structure induction. This is surprising, since as argued, time is a ubiquitous feature of human interactions with their environment, and the need to process rich temporal information in real time is a practical constraint on most of our basic causal inferences. In this paper we take a rational analysis approach (Anderson, 1990; Simon, 1982), starting with a normative account of inference from observations of real-time events to their underlying causal structure and developing a process-level approximation family that can capture human deviations from this. For our normative account, we expand the CBNs framework so that it incorporates representing and learning via causal delay information. Alongside this, we propose a process-level framework that exploits several tricks for approximating intractable probabilistic inference: mental simulation (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018), local computations (Bramley, Dayan et al., 2017; Fernbach & Sloman, 2009), and temporally local evidence (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Bramley, Dayan et al., 2017; Bramley, Lagnado, & Speekenbrink, 2015).

## 1. Question 1: How do beliefs about causal orders and delays shape causal structure learning?

One of our main goals is to test whether people can use their knowledge about time and causality to learn causal structure. Previous studies have demonstrated the temporal knowledge from three perspectives: order, delay expectation, and delay variation.

Foundational to the notion of causation, is the principle that causes must precede their effects (Hume, 1740). Accordingly, people use the *order* of occurrence to constrain and sometimes fully attribute causal structure among components of a system (Bramley, Gerstenberg, & Lagnado, 2014). Indeed, event order appears to be a strong heuristic cue to causal order, having been shown to override contingency patterns even in settings where participants are instructed that order is an unreliable guide (Lagnado & Sloman, 2006) or even completely irrelevant to causal structure (Rottman & Keil, 2012).

As well as order, causal inferences are sensitive to *delays* between events. People make stronger or more confident (generative) causal attributions connecting events separated by short temporal delays than by long temporal delays (Shanks & Dickinson, 1991; Shanks, Pearson, & Dickinson, 1989; Tarpy & Sawabini, 1974). This reflects one of the most basic forms of learning, in which animals associate stimuli at a learning rate inversely related to their separation in time (Grice, 1948). However, going beyond automatic associations in time, human causal attributions are moderated by domain-specific delay expectations, with shorter-than-expected delays also reducing the causal judgment strength (Buehner & May, 2002; Buehner & McGregor, 2006; Hagmayer & Waldmann, 2002; Lagnado & Speekenbrink, 2010; Mendelson & Shultz, 1976). For example, Hagmayer and Waldmann (2002) found participants judged whether an insecticide prevents mosquitoes by comparing prevalence of mosquitoes in fields with and without the insecticide, but judged whether planting flowers prevents mosquitoes based on whether the prevalence of mosquitoes was affected the year after the flowers were planted, presumably expecting that flowers would take longer to influence the insect population than insecticide.

[1] This is a real life example for one of the authors of this paper.

Besides the length of inter-event delays, people are also sensitive to *delay variability* when they are repeatedly exposed to putative cause–effect pairs. That is, people rate one kind of event as less of a strong cause of another to the extent that the delay varies a lot across instances (Greville & Buehner, 2010; Lagnado & Speekenbrink, 2010).

Recently several studies proposed models to capture human's expectations for delay length and variation, including scenarios of pairwise causal learning (Pacer & Griffiths, 2012), structure learning (Bramley, Gerstenberg, Mayrhofer et al., 2018; Pacer & Griffiths, 2015), imputing hidden causes (Valentin, Bramley, & Lucas, 2022), or making judgments of actual causation given a known causal structure (Stephan, Mayrhofer, & Waldmann, 2020). Nevertheless, these studies have predominantly focused on cases of generative causal influence. Additionally, they have focused on inference from sets of independent clips, in which root components are usually activated at the start and effects follow from this. However, a more naturalistic and challenging setting is one where causes and effects intermingle and components can exhibit multiple activations, and both generative and preventative influences can occur within a single learning episode. This is the setting we will explore.

## 2. Question 2: How do generation, prevention, and background causes interact in affecting causal learning?

Early studies of causal cognition focused on elemental pairwise causal judgments based on contingency data. While not directly related to the current temporal setting, these studies reveal general principles of causal inference. For instance, the $\Delta P$ principle captures the change in the probability of an effect occurrence with vs. without a putative cause ($P(E|C) - P(E|\neg C)$), forming a basic metric for the strength and direction of a potential causal effect (Allan, 1980). However, researchers later found people are sensitive to the *base rate* of the effect $P(E|\neg C)$. That is, how frequently the effect occurs in the absence of the cause. For a fixed $\Delta P$, people infer stronger generative influences when base rates are high (because this implies the cause would have succeeded a greater proportion of the time if it had the chance to operate), and stronger preventative influences when base rates are low (Buehner et al., 2003; Cheng, 1997; Wu & Cheng, 1999).

In addition to the size of the base rate, the regularity of the base rate also influences causal inference. In Rottman (2016), participants were asked to evaluate the effectiveness of two medications. In one context, the baseline pain level was random from case to case, whereas in another setting, it was autocorrelated (i.e. it tended to increase or decrease smoothly over time). Participants were found to focus more on the raw effect values in the random condition, while focusing more on the change of effect values in the autocorrelated condition. This indicates that people are sensitive to environmental *stability*, adapting how they accumulate and represent causal effect evidence when receiving information in different environments (Biele, Erev, & Ert, 2009; Whittle, 1988). We will explore whether people are sensitive to temporal regularity (periodic vs. unpredictable) and, if so, whether or not they adjust their inference strategy accordingly.

Finally, humans show some ability to condition on other variables when inferring the role of a target variable (Beckers, De Houwer, Pineno, & Miller, 2005; Gopnik, Sobel, Schulz, & Glymour, 2001; Rescorla & Wagner, 1972; Shanks, 1985). People can use information regarding known causes to better understand unknown causes, particularly preventative causes. The classic paradigm in prevention learning is to let learners build a generative impression of a cause ($A+$), and then expose them to negative results under the combination of a generative cause and a preventative cause ($AB-$). People learn the preventive cause better in this case than when the preventative cause is paired with the negative result alone ($B-$, Lee & Lovibond, 2021; Lovibond & Lee, 2021; Melchers, Wolff, & Lachnit, 2006; Rescorla & Wagner, 1972). However, the existence of temporal information may actually increase the difficulty of thinking about causal interactions: To utilize the generative causes to learn about prevention, the learner must have ensured that generative causes would have produced effects in a particular time period when preventative causes are active.

Recent studies also demonstrate human limitations in dealing globally with joint probability, i.e. reasoning probabilistically about multiple interacting variables (Bonawitz et al., 2014; Davis, Bramley, & Rehder, 2020; Fernbach & Sloman, 2009; Griffiths, Lieder, & Goodman, 2015; Markant, Settles, & Gureckis, 2016). Outside of very simple learning problems, they may rather focus on local components of the system rather than maintain a global perspective. For example, people often infer an erroneous $A \rightarrow C$ link when reasoning about a generative system with two links $A \rightarrow B \rightarrow C$, apparently failing to notice that $B$ can explain $C$'s dependence on $A$ (Davis et al., 2020; Fernbach & Sloman, 2009). Through model comparison, we will explore to what extent people can reason globally or locally about causal structure on the basis of real time evidence, e.g. whether they can account for and potentially bootstrap their inferences by considering interactions between causal mechanisms, or if they rather fail to make these accommodations.

## 3. Question 3: How do people process temporal dynamics to make causal inferences?

We build two models for describing how the temporal information could be processed in order to make causal inferences. We will explain the models at a theoretical level in this section and refer the readers to Appendices B and C for technical details. To do this, we first introduce the learning task before describing our model so that readers can get a concrete understanding of how it works.

### 3.1. The learning task

In this study, participants must guess the structure of abstract causal "devices" (Bramley, Dayan et al., 2017; Bramley, Gerstenberg, Mayrhofer et al., 2018; Gong, Gerstenberg, Mayrhofer, & Bramley, 2023) composed of three components (Fig. 1a–d): two "control components" (i.e. Cause $A$, $B$) and one "target component" (i.e. Effect $E$) on the basis of observations of those structures being perturbed by interventions. To control the impact of interventions, our experiments focus on a learning setting wherein the interventions are part of the stimuli, meaning participants observe them taking place rather than selecting and performing them themselves. For each device, the connection between each control component and the target component could be generative, preventative, or they might be unconnected (non-causal). Thus, we focus on learning in a nominal hypothesis space of 9 possible structures including all combinations of generative, preventative and non-causal connections from $A$ and $B$ to $E$ (Fig. 1e). As a first foray into preventative causation in real-time causal structure induction, we focus on this restricted hypothesis space of causal structures which only contains the common effect topology. However, the experimental paradigm and computational models we introduce generalize directly to learning in arbitrarily broader causal hypothesis spaces, as well as under different prior expectations about plausible delays and relations.

We focus on relationships between *point events* (i.e. activations) occurring at a device's components at particular moments in time. We assume an activation of a generative component will always produce an "extra" activation of the target component (i.e. causal strength = 1, Cheng, 1997, see Fig. 2a). We use the gamma distribution to model and generate the delays between causes and effects (Bramley, Gerstenberg, Mayrhofer et al., 2018; Stephan et al., 2020; Valentin et al., 2022). See Appendix A for more details.

We assume an activation of a preventative component blocks any activations of the target component for a short stochastic time window (Fig. 2b). We assume that prevention occurs irrespective of whether activations would have been caused by a generative causal influence or would have occurred spontaneously. Preventative influences are thus conceived as having a broad preventative scope (Carroll & Cheng,

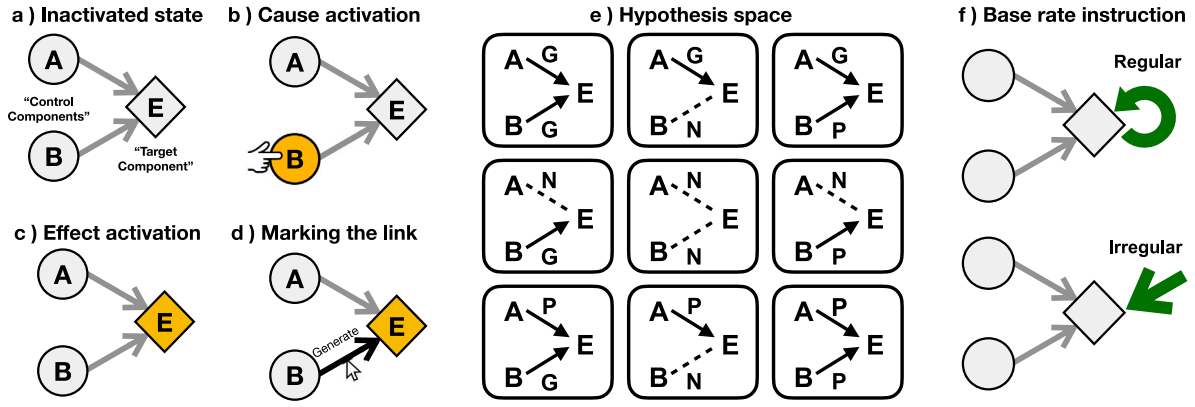


**Fig. 1.** Causal devices tested in this paper. (a–d) Experimental interfaces. Participants were instructed to the control components and target components in the causal devices and observed how the system reacted to pre-set interventions. They marked their answers of the role of each connection during or after the observation. (e) The response hypothesis space (all possible pairwise combinations of generative (G), non-causal (N), and preventative (P) connections). (f) The illustrations shown to participants in the regular (periodic) vs. irregular (exogenous) base rate condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

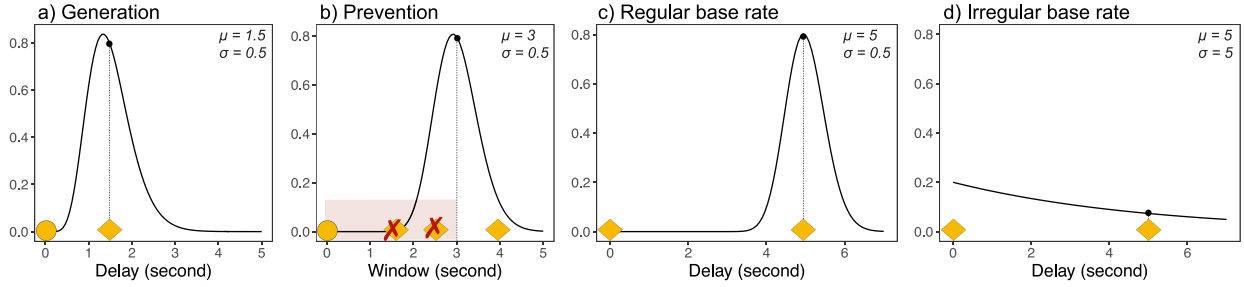


**Fig. 2.** Using gamma density distributions to generate the delays between cause and effect and the blocking windows of preventative causes. Circles indicate cause events and diamonds indicate effect events. Each vertical line shows an actual sampled situation. (a) The distribution of delays between cause and effect. When a generative cause event occurs, it will produce an effect event after $1.5 \pm 0.5$ s. (b) The distribution for preventative window length. When a preventative cause event occurs, all effect events supposed to occur within $3 \pm 0.5$ s will be canceled, while effects outside the preventative window (the red box) would not be affected. (c) The distribution of delays between base rate events in the regular condition. When a base rate effect occurs, the next base rate effect will occur after $5 \pm 0.5$ s. (d) The distribution of delays between base rate events in the irregular condition. When a base rate effect occurs, the next base rate effect will occur after $5 \pm 5$ s. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2009).[2] By definition, activations of non-causal components have no impact on the behavior of the target component.

Two forms of background activation are considered. In the *Regular base rate* condition, the target component activates quasi-periodically (Fig. 2c). In the *Irregular base rate* condition it occurs exactly as often overall but is completely unpredictable when the next occurrence will be (Fig. 2d; see Appendix A).

### 3.2. Bayesian inference

We now lay out an ideal Bayesian model as a normative model for this task. The ideal reasoner is presumed to take all activation events within the observation interval as the basis of their inference and use the relative likelihood of these under different structural hypotheses to update a distribution over causal structures. The calculation of likelihood here depends on an expensive enumerative actual causal attribution step (Halpern, 2016). The basic idea is that accurate judgments about *type-level* causal relationships (i.e. about the underlying causal structure) depend on detailed considerations about the *token-level* causation giving rise to the observable evidence (i.e. which particular event actually caused which particular effect). There are often a very large number of possible ways that even a single causal hypothesis could have produced a particular pattern of observed events. For instance, if $A$ activates at 0.1 s and $B$ activates at 1.2 s ($\mathbf{i}\{i_A^{(1)} = 0.1s, i_B^{(1)} = 1.2s\}$), and the learner observes two subsequent effects ($\mathbf{d}\{d^{(1)} = 1.5s, d^{(2)} = 2.8s\}$), even under the hypothesis that $A$ and $B$ are both generative causes, the data could be produced in multiple ways: $A$ could have caused the first effect and $B$ the later one ($i_A^{(1)} \rightarrow d^{(1)}$, $i_B^{(1)} \rightarrow d^{(2)}$) or $A$ could have caused the later effect and $B$ the earlier one ($i_A^{(1)} \rightarrow d^{(2)}$, $i_B^{(1)} \rightarrow d^{(1)}$). Alternatively one or both connections could have not revealed their effects yet and meaning either or both observed effects could simply be base rate activations. Therefore, in order to maintain rational beliefs about causal structure, the ideal reasoner considers all possible causal paths that could describe what actually happened given each possible structural hypothesis.

Fig. 3a shows two examples of the tree of possible causal paths under two of the possible structural hypotheses. Since one must consider possible causal paths exhaustively, the complexity of this inference scheme scales in a worse-than-polynomial manner as the number of events a learner observes increases.

### 3.3. Simulation-and-summary-statistic approximation

While the enumerative approach achieves benchmark performance by inverting the generative model, exhaustively considering pathways linking all observed events, it makes unrealistic demands on memory storage and computing power compared to what could plausibly be involved in human cognition. Therefore, we propose a process-level model that is more consistent with cognitive constraints. It is based on the simulation-and-summary-statistic idea (also written as "summary-statistic" for short), which is an important approach in Approximate

[2] We recognize that there are other ways in which one might operationalize prevention and we consider several alternatives in the General Discussion.

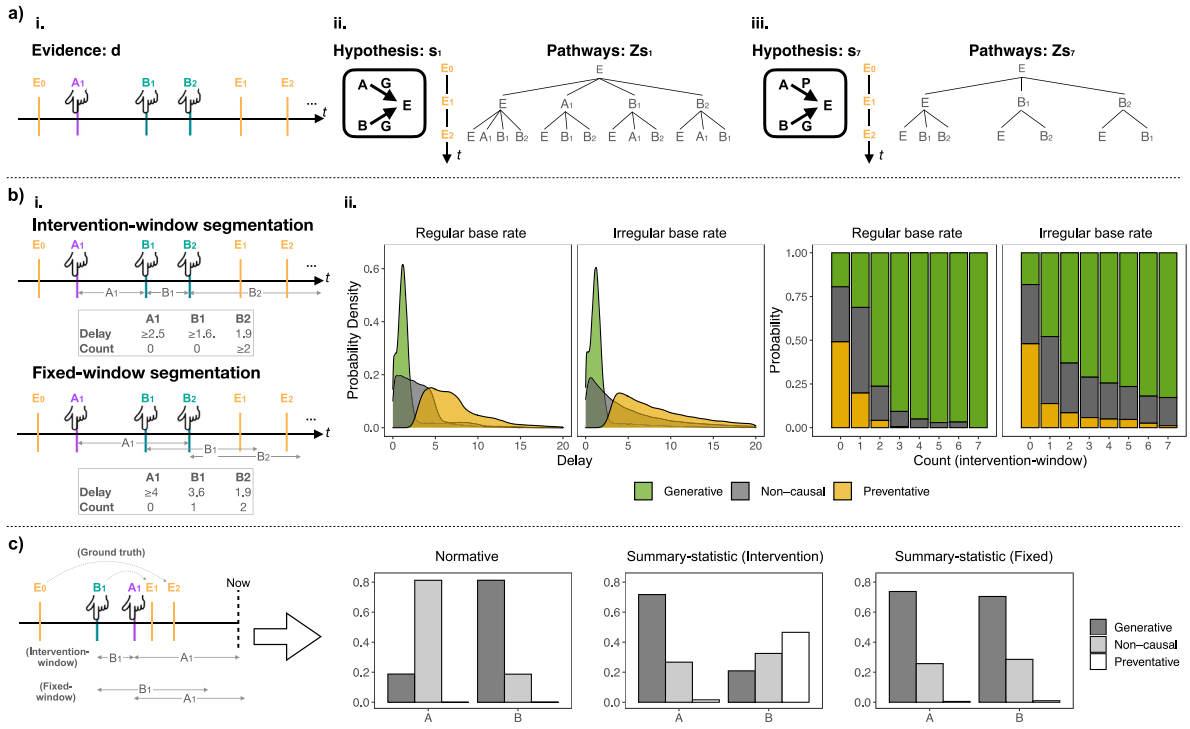


**Fig. 3.** Illustrations of model algorithms. (a) Causal path construction under fully normative inference. i. Data: Each line indicates a cause ($A$, $B$) or an effect ($E$) event in the evidence. ii-iii. Ideal observer sums over all possible pathways (branches) that explain all events evidence under each hypothetical structure. ii. e.g. Under the structure where $A$ and $B$ are both generative causes, there are 13 ways to explain Evidence $d$: one candidate cause for $E_0$ (base rate), four candidate causes for $E_1$, and 3–4 candidate causes for $E_2$ depending on how $E_1$ is explained. iii. Possible pathways under a different structure. (b) Summary-statistic approach: i. Intervention-window or fixed-window evidence segmentation. ii. Distributions for summary-statistics given different connection types based on pre-simulated data. The model uses likelihood of observed statistics under these distributions as a proxy for generative model likelihood. Distributions slightly differ given different base rate conditions. (c) Example where posterior over structures differs among models (assuming a regular base rate). Curved arrows indicate the true underlying generative process unknown to the models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Bayesian Computation in statistics (Blum, Nunes, Prangle, & Sisson, 2013; Lintusaari, Gutmann, Dutta, Kaski, & Corander, 2017; Sunnåker et al., 2013; Zhao et al., 2023). We explore this idea's cognitive plausibility as an explanation for human judgments in our setting. Our model incorporates three features of bounded inference that are often highlighted in cognitive psychology: mental simulation, local computation, and temporally local evidence.

### 3.3.1. *Mental simulation*

The tendency to rely on simulation-based approximation to exact inference has been hypothesized to play an important role in model-based reasoning in many scenarios, including physical scene understanding (Battaglia et al., 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Ullman et al., 2018), mechanical reasoning (Hegarty, 2004), and causal judgment (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). The idea is that instead of computing the likelihood of a potential generative model producing observed data exactly, people instead compare their observations to mental simulations of what *kind of pattern* they expect to happen under different generative models.

Critical to this process is the identification of a useful set of easily tracked abstract cues or features with which to compare such simulations to observations. When a scenario of interest involves complex dynamics, direct surface-level (i.e. "pixel-level") comparison between simulated and observed evidence is generally inappropriate for measuring the likelihood of a hypothesis. Ullman et al. (2018) combined the ideas of simulation and abstraction to model inferences about the latent properties of physical objects (such as masses and forces) from observed dynamics. As a simple example, if imagined heavy objects tend to move more slowly than imagined light objects, this licenses the use of speed as a (fallible) cue to mass.

Concretely, we explore whether simple salient local features of event sequences that are diagnostic (if fallible) guides to local causal relationships can explain human judgments better than a fully Bayesian treatment. The implied cognitive process is that learners draw on (imagined) evidence under different causal ground truth structures in order to develop statistical cues that can be directly applied to pairwise causal judgments. Here we simply investigate two straightforward and salient cues that people might be sensitive to in the current task:

1. **Delay:** The interval between a cause component's activation and the next subsequent effect activation.
2. **Count**: The number of activations of the effect after the cause activation within some time window.

These cues are hand-engineered, and far from exhaustive. However, they are simple to track and turn out to discriminate reasonably well between different types of causal connections. As shown in Fig. 3b, for the delay cue, we generally expect to see shorter intervals between a control component's activation and the target component's next activation if the control component is a generative cause, a medium and more variable interval if there is no connection or a longer interval if it is preventative. For the count cue, more effect activations are likely to follow the activation of the generative component on average because of the existence of base rate activations as well as generation. In contrast, fewer activations are likely to follow the activation of preventative components. The former cue considers concrete delay information but ignores the possibility of different causal pathways, while the latter cue ignores the exact temporal interval between events (cf. Bramley, Gerstenberg, Mayrhofer et al., 2018).

### 3.3.2. *(Structurally) local computation*

Both the count and delay cues introduced above ignore surrounding structure and context leading to the potential for interference. For

example, in the presence of a known preventative cause that has just occurred, an ideal learner should reduce their expectation that a generative cause would produce a short delay to the next event, or a high subsequent effect count. Thus, this approach also captures a principle of local computation (Bonawitz et al., 2014; Bramley, Dayan et al., 2017; Davis et al., 2020; Fernbach & Sloman, 2009; Griffiths et al., 2015; Markant et al., 2016), predicting that learners will make causal attributions at the level of individual links without accommodating the global context and the full space of global causal models.

The other reason why we apply local computation to this process-level model is that it can greatly reduce the computational cost compared to the global computation approach. In the current continuous-time setting, interventions could happen at any time making every context unique. This means that conditioning one's inference on even a single previous intervention requires learners to simulate a much larger number of one-off context-specific situations. Introducing more of this context sensitivity (i.e. constructing separate summary statistics for each possible combination of causes) would allow a summary-statistic approach to perform closer to normative inference but at the cost of increasing computational demands and reducing generality beyond the set of contexts considered.

#### 3.3.3. (Temporally) local evidence

The final cognitive feature we consider is related to how people parse and segment the evidence encountered across an extended observation of a causal system. Ullman et al. (2018) applied a summary-statistic approach to short observations (5 s) and allowed participants unlimited replay opportunities, so assumed people could use cues based on the entire observation. In the scenario considered here, the learner observes causal dynamics for considerably longer (20 s, containing dozens of events) without recourse to replays. In general, we experience the world in a single ongoing timeline. Thus, with finite short-term memory storage and attention, it seems plausible that people abstract cues more locally than from full observation. In other studies, people are found to often use temporally local (i.e. recent) information to drive causal model learning (Bramley, Dayan et al., 2017; Davis et al., 2020; Rehder, Davis, & Bramley, 2022). Furthermore, people are often unable to recall older evidence exactly (Bramley et al., 2015; Harman, 1986), rather remember whatever conclusions they have drawn on the basis of it.

In line with these ideas, we hypothesize that people segment their observations as they unfold, using recent events to update their beliefs and then discarding their memory of them. We consider two ways to segment continuous-time evidence. As shown in Fig. 3b, a unit of evidence under both approaches begins with an intervention (i.e. the activation of a control component), capturing the basic principle that causes can only influence what happens later. An *Intervention-window* segmentation approach treats one unit of observation as the interval between one intervention and the next. This removes the distraction of other interventions that might also influence the effect, but ignores the fact that these interventions might be performed irregularly or reactively, and also that actual effects may not have been revealed before the occurrence of the next intervention. A *Fixed-window* approach ends one unit of observation after a fixed amount of time. This has the advantage of stability in its odds of including all relevant effects' but instead opens the door to confounding influences when subsequent interventions occur within the preceding observation window. A fixed window approach also implies some degree of parallel processing since fixed-length attentional windows may easily overlap in a single timeline.

### 3.4. Summary of modeling frameworks

In sum, we have laid out two approaches to solving the current learning problem. The normative model utilizes the exact timing information of each event, considering all possible observation-consistent ways in which the effects might have been generated or prevented. The summary-statistic model compresses the information by abstracting useful cues and comparing the similarity between cues summarized from observation with mental simulation. We do not see the two accounts as fundamentally in tension. Rather, the summary-statistic approach embodies a set of algorithmically plausible steps to approximate the normative solution.

Given the information compression and the local focus of the summary-statistic approach, its predictions diverge from the normative one in some situations. One example is shown in Fig. 3c. When $B$ activates and then $A$ activates followed closely by two effects, the normative learner finds this most consistent with the structure where $B$ is a generative cause because the delay between $B$ and the first effect is consistent with its delay expectation, while the other effect could easily be due to the base rate. For the summary-statistic models, the intervention-window approach suffers from a blocking effect, where the occurrence of $A$ masks any potential link between $B$ and the effects. The fixed-window approach suffers from a local computations error, where each effect is potentially attributed to both $A$ and $B$ leading to a marginal preference for the model with both $A$ and $B$ as generative causes. We will show more similarities and differences between the two modeling approaches alongside human behavior in Results sections.

## 4. Overview of experiments

We now report on three experiments that investigate how people infer preventative and generative causal structures in continuous time. Each experiment includes stimuli generated from each of the nine underlying structures we consider (Fig. 1e). Experiment 1a and 1b aimed at exploring how overall structure and regular and irregular base rates influence causal judgments. Experiment 2 additionally includes stimuli designed to probe whether people make specific mistakes predicted by the summary-statistics model. All pre-registrations, materials, data, and analysis code are available at https://osf.io/q8n72/. Stimuli for all experiments can be viewed at https://github.com/tianweigong/causal_diamond.

## 5. Experiment 1

### 5.1. Methods

#### 5.1.1. Participants

One hundred and eighty-seven participants from Amazon Mechanical Turk were recruited and reported for Experiment 1a (81 female, 105 male, 1 non-binary, aged $37 \pm 11$, regular vs. irregular condition: 93 vs. 94) and another 123 participants were recruited and reported for Experiment 1b (45 female, 78 male, aged $39 \pm 11$, regular vs. irregular: 63 vs. 60). The sample size of Experiment 1a was determined by a power analysis comparing two between-subject groups anticipating a medium sized effect ($d = 0.5$) with a goal of .90 power at the standard .05 alpha. The sample size for Experiment 1b followed a pilot study (Gong & Bramley, 2020) given that both of them aimed to compare participants' performance with normative and heuristic models. Nine additional participants in Experiment 1a were recruited but excluded prior to analysis because they clicked (to respond) more than 300 times during the task (as average participants acted 113 ± 26 times). Hence, we suspected these respondees were either inattentive or non-human. Four additional participants in Experiment 1b were recruited but excluded prior to analysis because they clicked more than 300 times during the task (n = 2), or failed to pass at least one of two attention questions (n = 2).[3] Participants were paid between $1.00 and $2.08

[3] We also pre-registered to exclude participants who took more than six attempts to pass all instruction comprehension check questions. However, with the benefit of hindsight, we recognized that even attentive participants often required several attempts to pass our stringent comprehension checks. Thus, we opted to relax this exclusion criteria.

depending on their performance (see below) and experiments lasted around 20 min.

#### 5.1.2. Design & procedure

*Overview.* In both Experiment 1a and 1b, participants judged the causal structure of 18 causal devices (Fig. 1e). When a generative cause event occurred, it would produce an effect event after $1.5 \pm 0.5$ s (see Fig. 2a). Whenever a preventative cause event occurred, any upcoming effect events in the subsequent $3 \pm 0.5$ s were canceled (see Fig. 2b). Each base rate event occurred $5 \pm 0.5$ s after the previous one in the regular base rate condition, or $5 \pm 5$ s in the irregular base rate condition. The choice of generative delay was based on past studies that suggest people only reliably attribute causal relations to delays of up to around 2 s in the absence of context information shaping delay expectations (Shanks & Dickinson, 1991; Shanks et al., 1989). We chose the size of the true preventative windows and base rates such that base rates are generally lower than casual influences (i.e. activity is relatively sparse without any generative events) and preventive influences last long enough to have a reasonable chance of preventing something. The true sampled causal delays are unknown to the learner (human or model), but for simplicity we pre-trained (Experiment 1a) or told (Experiment 1b) participants about typical patterns of base rate activations and about typical generative delays and preventative durations in an instruction phase, and so also assumed these parameters were available to all models.

For each device, participants clicked a "Start" button to watch the clip. Each clip started with a base rate activation of the target component and included three pre-set interventions on $A$ and three on $B$ randomly spaced and intermingled over 20 s. After that, the clip would end and no further activations could be observed. Components' activations were displayed as the component "lighting up" by changing from gray to yellow for 350 ms. The activation of the control component was accompanied by a hand symbol (Fig. 1b) and participants were told that this showed that control components were being intervened on by someone or something external to the system, meaning that the interventions happened at random moments rather than following any informative pattern. Clips were selected to make sure that no activation was masked by another on the same component in the clips, and participants were also told about this rule.

Participants were invited to mark their guesses about the two connections during or after the clip by clicking the space between the components (Fig. 1d). Each clip could only be played once. The order of 18 trials, as well as the click pattern (whether they would have to click once, twice, or tree times to select generative, preventative or non-causal), and the vertical position of $A$ and $B$ components (above or below) were randomized independently between participants.

Participants were informed of the timing of three types of connections as well as the target component's self-activation prior to the inference task. For the base rate specifically, participants in the regular condition were told that the target component would activate regularly about every five seconds and they saw an illustration with a circular arrow to create the impression of periodic activation (Fig. 1f). Participants in the irregular condition were told that the target component can activate by itself at completely random times and they saw an illustration with an exogenous link intended to imply that someone sometimes activates the target component directly but one cannot anticipate when it will happen (Fig. 1f). In order to similarly provide timing information, participants were told the base rate activation happens about 2–7 times per clip. Participants had to pass introduction check questions before starting the experiment. To properly incentivize accurate judgments, a 3-cent bonus was paid for each correctly identified connection and non-connection during the main task in addition to the basic $1 payment.

*Experiment 1a.* In Experiment 1a, to generate stimuli from different structures (e.g. both generative, one generative and one non-causal) and different conditions (i.e. regular vs. irregular) comparable, we used a Latin-square design. We first created 18 causal delay seeds independently. Each of these included a set of timings for interventions, base rate activations, which depended also on whether the base rate was regular or irregular, and what generative delays (or blocking windows) $A$ and $B$ would have if they were generative (or preventative) components. Under each seed, 18 stimuli (9 causal structures $\times$ 2 base rate settings) were generated by implementing generative or preventative influences according to the ground-truth structure (see Fig. 4a for an example of a single seed manifesting under each structure and base rate condition). Across different seeds, the timing and order of interventions were randomly generated to capture the diversity of ways in which the interventions could be interleaved ranging from perfectly interleaved (e.g. $ABABAB$) to perfectly clustered (e.g. $AAABBB$, see Fig. 4b for an example of a single structure under different seeds and base rate conditions). All stimuli were finally divided into 18 sets (9 sets for each base rate setting) according to a Latin-square design that ensured participants would only see one structure under each seed (see https://osf.io/sqv6c for the counterbalancing matrix). Participants were randomly assigned to one of these 18 sets.

In the instructions, participants saw training videos that showed the patterns of the target component's base rate activations (corresponding to their condition) and also what happens after intervening on a causal system with a single (generative, non-causal, or preventative) connection. They completed a single practice trial in which the true causal device included one generative connection and one non-causal connection. Feedback was provided in the practice trial but not in the test trials.

*Experiment 1b.* Experiment 1b differed from Experiment 1a from two perspectives. Firstly, although we assume the provenance of the summary-statistic approximation to be mental simulation, cues might also be derived from experience with the "labeled data" included in the instructions or practice trials. Therefore, Experiment 1b only kept the text instructions and removed the training videos and practice trials, to show that labeled data were not necessary for participants to complete the task.

Additionally, given that the stimuli in Experiment 1a were generated by one of the ground truth structures, the normative model and summary-statistic approximations often made similar predictions. To probe how participants react to situations with stronger discrepancies between the normative and summary-statistic predictions, we created some stimuli that were not generated by any particular causal device. We created two blocks of stimuli in Experiment 1b. Block 1 included nine stimuli for each participant, which replicated the procedure of Experiment 1a, and served to ensure participants were habituated to reacting to "normal" stimuli. In Block 2, we generated potential test stimuli by randomly distributing six interventions and between 1 and 9 effects across a 20 s trial. We selected sequences for which the structure predictions of the normative and summary-statistic models were strongly dissimilar, while ensuring that these stimuli were not too normatively improbable (i.e. that they could conceivably have been generated by one of the causal structures).[4] There were 27 stimuli for each condition and each participant observed nine of them. Block 1

[4] Specifically, we picked the stimuli where at least one (intervention-window) summary-statistic cue (Delay or Count) had a different dominant answer compared to the normative model and rejected any for which the likelihood of the most probable structure producing the data was extremely low ($<10^{-40}$). The squared error between normative and summary-statistic predictions in Block 1 (trials with the ground truth) and Block 2 (trials without the ground truth) was 0.22 vs. 0.53 on average. The likelihood of the most probable structure according to the normative model in Block 1 and Block 2 was 0.07 vs. 0.004 on average.

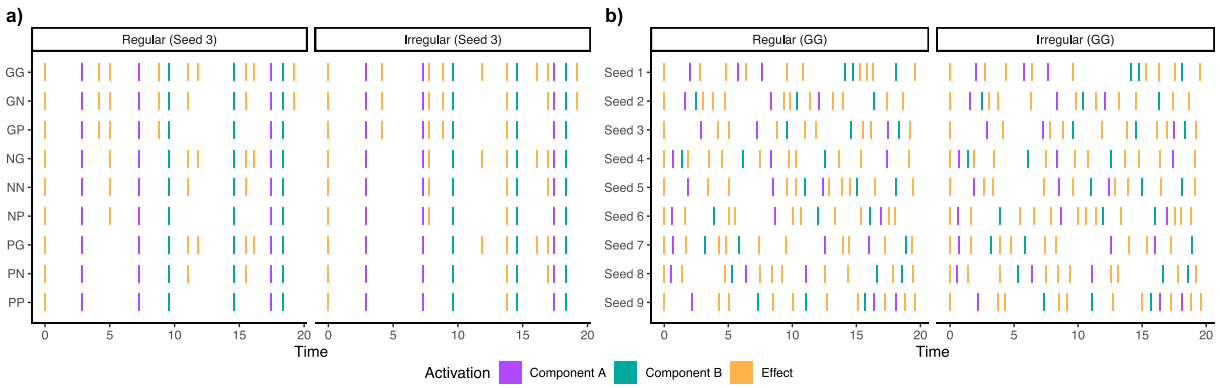


**Fig. 4.** (a) Examples of a single seed under different structures and base rate conditions (from one stimulus seed used in Experiment 1a). *Y*-axis refers to the roles of Component *A* and *B* (e.g. GP: *A* is a generative cause and *B* is a preventative cause). (b) Examples of a single structure manifesting under different seeds and base rate conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

always preceded Block 2 so that the first half task would be identical to Experiment 1a. Participants completed 18 trials in sequence without any delineation between the blocks. All other experimental settings remained identical to Experiment 2. The bonuses were, in reality, determined by doubling the bonuses participants gained in Block 1.

### 5.2. Results

We focus on analyzing participants' accuracy by comparing their judgments against the ground truth. We investigate whether participants' performance was influenced by the nature of the underlying causal mechanism, base rate regularity, or the observed intervention sequence (i.e. whether this involves interleaved interventions on the two components or clusters of interventions on one component then the other). Since these analyses require there to be a correct answer, for Experiment 1b we only include Block 1.

To compare our models' behavior qualitatively with participants', we simulate judgments of each model type after observing the same stimuli as the participants. We used one fitted softmax parameter for each model and repeated each simulation 200 times per participant to obtain stable and consistent distributions of simulated judgments (see Appendix D for model fitting details). For summary-statistic models, we average predictions under the two proposed features with equal weights to form a combined prediction (cf. Ullman et al., 2018). Results of intervention-window vs. fixed-window summary-statistics were similar at the aggregated level, and hence we only visualize the intervention-window results in the figures.

#### 5.2.1. Overall performance

In Experiment 1a, participants in both regular and irregular conditions performed above chance at the connection level (chance = 33%, regular: 66% ± 22%, $t(92) = 14.75$, $p < .001$, $d = 1.53$, 95%CI of $d = [1.23, 1.84]$; irregular: 61% ± 18%, $t(93) = 14.67$, $p < .001$, $d = 1.52$, 95%CI of $d = [1.22, 1.82]$) as well as the structure level (1 = correct in both connections; 0 = otherwise, chance = 11%, regular: 49% ± 27%, $t(92) = 13.83$, $p < .001$, $d = 1.43$, 95%CI of $d = [1.15, 1.73]$; irregular: 41% ± 22%, $t(93) = 13.27$, $p < .001$, $d = 1.37$, 95%CI of $d = [1.09, 1.66]$). These patterns were replicated in Experiment 1b, where participants also performed above chance at both connection (regular: 67% ± 22%, $t(62) = 11.93$, $p < .001$, $d = 1.50$, 95%CI of $d = [1.15, 1.88]$; irregular: 59% ± 19%, $t(59) = 10.11$, $p < .001$, $d = 1.30$, 95%CI of $d = [0.96, 1.66]$) and structure levels (regular: 49% ± 29%, $t(62) = 10.64$, $p < .001$, $d = 1.34$, 95%CI of $d = [1.00, 1.69]$; irregular: 39% ± 23%, $t(59) = 9.21$, $p < .001$, $d = 1.19$, 95%CI of $d = [0.86, 1.53]$). Indeed, accuracy did not differ between Experiment 1a and 1b at the connection level (regular: $t(154) = 0.09$, $p = .926$; irregular: $t(152) = 0.81$, $p = .418$) or the structure level (regular: $t(154) = 0.004$, $p = .997$; irregular: $t(152) = 0.56$, $p = .578$). This means that labeled data in the form of video training and practice trials were not a necessary condition for participants' success in this task. We therefore combine stimuli from two experiments in later analyses to obtain a larger sample size.

#### 5.2.2. Focal and neighboring causes

To investigate participants' ability to identify generative, non-causal, and preventative connections, as well as whether the base rate regularity or the neighboring connections would influence performance, we performed a 3 (focal cause: generative, non-causal, preventative) × 3 (neighboring cause: generative, non-causal, preventative) × 2 (base rate regularity: regular, irregular) mixed ANOVA. Each trial provided two data points here, one regarding *A* as the focal cause and *B* as the neighboring cause and the other regarding *B* as the focal cause and *A* as the neighboring cause.

There was a main effect of focal cause ($F(2, 616) = 101.24$, $p < .001$, $\eta_p^2 = .247$, 95%CI of $\eta_p^2 = [.200, .293]$). Participants performed best at identifying generative connections (77% ± 24%), then preventative connections (63% ± 31%), and finally non-causal connections (51% ± 29%, Fig. 5a). The differences were all pairwise-significant (Bonferroni adjusted $p < .001$).[5]

There was a main effect of base rate regularity ($F(1, 308) = 7.07$, $p = .008$, $\eta_p^2 = .022$, 95%CI of $\eta_p^2 = [.003, .057]$). Participants tended to perform better in the regular (66% ± 22%) than the irregular (60% ± 19%) condition. However, there was an interaction between focal cause and base rate regularity ($F(2, 616) = 3.69$, $p = .026$, $\eta_p^2 = 0.012$, 95%CI of $\eta_p^2 = [.001, .028]$). Analysis of the simple effects showed that the regularity difference was only significant for preventative causes (Fig. 5a). This is consistent with the principle that identifying preventative causes relies heavily on having a good counterfactual expectation of what would have happened in the causal system in the absence of the focal cause.

The main effect of neighboring cause was non-significant ($F(2, 616) = 2.76$, $p = .064$) while there was a interaction between neighboring cause and base rate regularity ($F(2, 616) = 6.66$, $p = .001$, $\eta_p^2 = .021$, 95%CI of $\eta_p^2 = [.005, .042]$). The neighboring connections made a difference in the irregular condition ($F(2, 308) = 8.56$, $p < .001$, $\eta_p^2 = .053$, 95%CI of $\eta_p^2 = [.017, .095]$), but not in the regular

[5] To rule out that this main effect was merely due to people generally selecting more answers as generative and preventative than non-causal, we calculated the F1-score for each cause (Powers, 2011). The patterns were the same when using the F1-score as the index ($F(2, 540) = 181.89$, $p < .001$, $\eta_p^2 = .403$, 95%CI of $\eta_p^2 = [.352, .448]$).

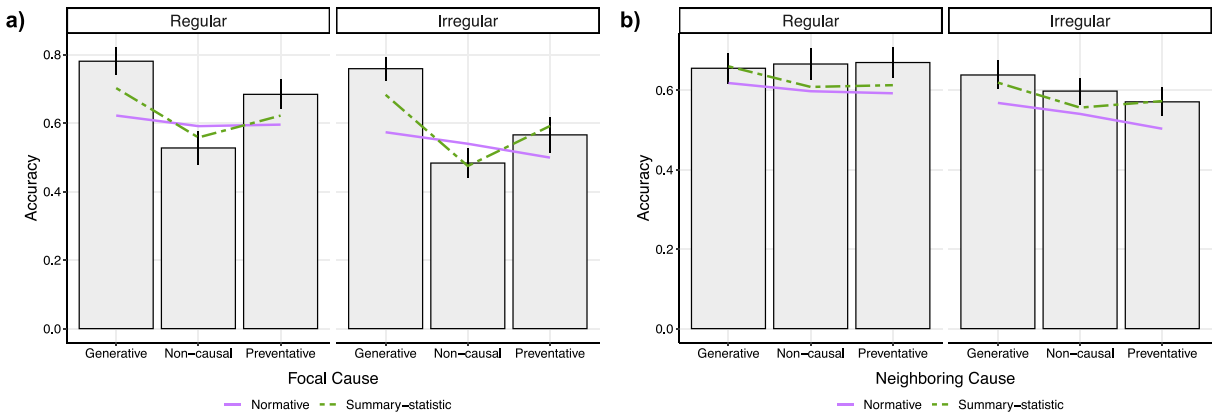


**Fig. 5.** (a) Accuracy of different causal connections in Experiment 1. (b) Accuracy in judging a connection (averaged across generative, preventative, or non-causal target connections) when paired with different types of connections in Experiment 1. Lines indicate the performance of simulated normative and summary-statistic learners each with a fitted softmax parameter based on participants' data in Experiment 1 (see Appendix D). Error bars indicate 95% confidence intervals.

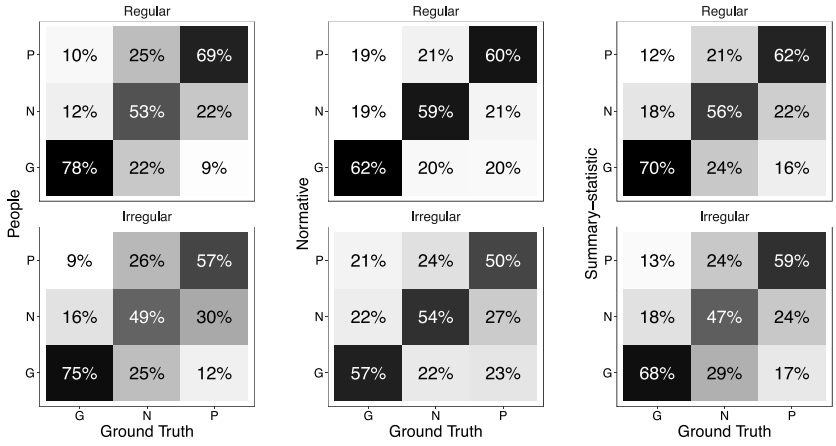


**Fig. 6.** Confusion matrices for participants' and models' choices under different ground truths in Experiment 1. The normative and summary-statistic learners were simulated with a fitted softmax parameter based on participants' data in Experiment 1 (see Appendix D).

condition ($F(2, 308) = 0.41$, $p = .662$, Fig. 5b). Participants in the irregular condition performed better when the neighboring connection was generative than non-causal ($t(308) = 2.48$, $p = .041$, $d = 0.099$, 95%CI of $d = [0.003, 0.195]$) or preventative ($t(308) = 4.13$, $p < .001$, $d = 0.165$, 95%CI of $d = [0.069, 0.261]$). This means that when the base rate was uncertain, a generative cause could stand in by setting up strong expectations. Other two-way or three-way interactions were non-significant ($ps > .05$).

For simulated model-based learners, the summary-statistic learner exhibited a similar tendency as participants, performing worse in identifying non-causal connections (Fig. 5a). The accuracy of both normative and summary-statistic learners was partly dependent on the neighboring cause. As shown in Fig. 3b, the summary-statistic distributions of the non-causal type, particularly the Delay distributions, frequently exhibit overlaps with both other distributions, and furthermore, the other types (generative or preventative) typically have higher density in the overlapping region.

### 5.2.3. Confusion matrices

Fig. 6 shows the proportion of participants' choices under different ground truths. We explored the frequency of choices when people made inconsistent judgments with the ground truth. Under the regular base rate, people were equally likely to judge a generative connection as a non-causal one or a preventative one (12% vs. 10%, chi-square goodness of fit: $\chi^2(1) = 3.01$, $p = .082$). They were equally likely to judge a non-causal connection as a generative or preventative one (22% vs. 25%, $\chi^2(1) = 2.65$, $p = .103$) while they more often judged a preventative connection as a non-causal one than a generative one (22% vs. 9%, $\chi^2(1) = 83.41$, $p < .001$). The results of irregular base rate were similar (non-causal ground truth: 25% vs. 26%, $\chi^2(1) = 0.70$, $p = .404$; preventative ground truth: 30% vs. 12%, $\chi^2(1) = 107.96$, $p < .001$) except now participants also more often judged a generative connection as non-causal than preventative (18% vs. 9%, $\chi^2(1) = 29.55$, $p < .001$). The summary-statistic learner exhibited a similar tendency to human participants, tending to mistake preventative or generative connections more often as non-causal, rather than mistaking one for the other.

### 5.2.4. Intervention order

We examined the influence of the intervention sequence. The intervention patterns in the experimental stimuli were randomly generated (albeit balanced to include 3 interventions per control component) and hence varied in terms of the sequence. In some trials, participants observed data in which interventions on one *component* were "interleaved" (e.g. $A$ in $ABABAB$ or $ABBABA$), in others they were fully "clustered" (e.g. $A$ in $AAABBB$ or $BAAABB$), and in others they were

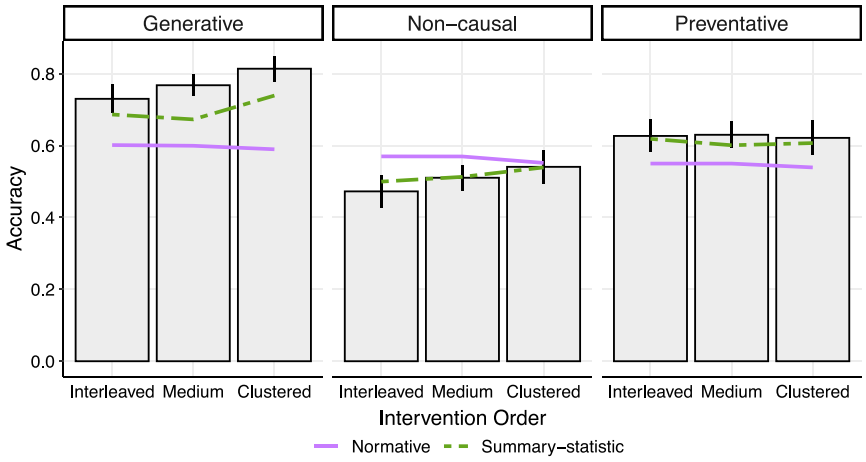


**Fig. 7.** Accuracy separated by intervention order in Experiment 1. Lines indicate the performance of simulated normative and summary-statistic learners each with a fitted softmax parameter based on participants' data in Experiment 1 (see Appendix D). Error bars indicate 95% confidence intervals.

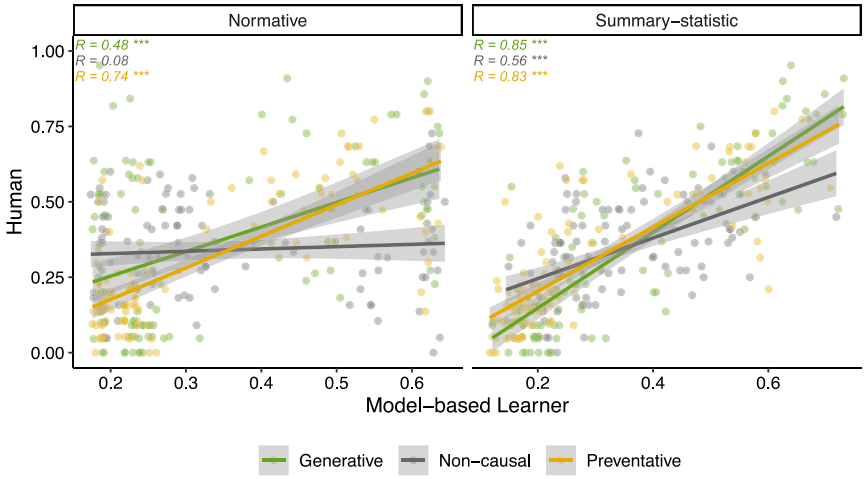


**Fig. 8.** Scatterplots of simulated model-based learners predictions and human judgments on the proportion of choosing different causal types in stimuli with no ground truth in Experiment 1b. Each connection in a stimulus is represented by three data points in the figure corresponding to the participant's and models' average probability assigned to that possibility. The normative and summary-statistic learners were simulated with a fitted softmax parameter based on participants' data in Experiment 1 (see Appendix D). Error bars indicate 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

partially clustered (e.g. $A$ in $AABABB$ or $ABBBAA$) which we called a "medium" level. We performed a 3 (focal cause: generative, non-causal, preventative) × 3 (intervention order: interleaved, medium, clustered) × 2 (base rate regularity: regular, irregular) mixed ANOVA. Each trial provided two data points, one regarding $A$ as the focal cause and the other regarding $B$ as the focal cause. The effects regarding focal cause and base rate regularity were similar to previous analyses and hence we only focus on the effects related to intervention order here.

There was a main effect of intervention order ($F(2, 538) = 9.39$, $p < .001$, $\eta_p^2 = .034$, 95%CI of $\eta_p^2 = [.012, .061]$) and an interaction effect between intervention order and focal cause ($F(4, 1076) = 3.22$, $p = .012$, $\eta_p^2 = .012$, 95%CI of $\eta_p^2 = [.001, .022]$). As shown in Fig. 7, the clustering intervention mainly benefited the identification of generative ($F(2, 269) = 11.40$, $p < .001$, $\eta_p^2 = .078$, 95%CI of $\eta_p^2 = [.031, .131]$) and non-causal ($F(2, 269) = 3.76$, $p = .025$, $\eta_p^2 = .027$, 95%CI of $\eta_p^2 = [.002, .063]$) connections, while the effect was insignificant for preventative connections ($F(2, 269) = 0.07$, $p = .935$). The summary-statistic learner demonstrated a similar influence of the intervention order as humans, while the normative learner performed indifferently across different intervention orders (Fig. 7).

#### 5.2.5. Trials optimized for model discrimination

Block 2 in Experiment 1b contained stimuli that were not generated from a particular ground truth structures, but rather generated so as to distinguish strongly between normative and summary-statistic models. Fig. 8 shows the choice proportion of human learners vs. simulated learners on each stimulus. The choices simulated from the summary-statistic model were better correlated with human judgments across generative, non-causal, and preventative answers. In particular, the summary-statistic model captured when people tended to judge a variable as non-causal (gray points and line) which often diverged from the normative prediction.

#### 5.2.6. Model fitting

To check quantitatively how well the models we have considered capture participants' causal judgments, we fit all participant judgments with our normative and summary-statistic models at both aggregate and individual levels. The details of the model fitting procedure can be found in Appendix D.

Participants choices were best captured by the summary-statistic approach, specifically by the variant that segments evidence according

**Table 1**
Model fits.

| | CV | BIC | $\tau$ | N (Regular) | N (Irregular) |
|---|---|---|---|---|---|
| *Experiment 1a* | | | | | |
| Normative | −6054 | 12 110 | 0.44 | 17%(14%) | 19%(13%) |
| SS (intervention-window) | **−5857** | **11 718** | 0.23 | **53%(47%)** | **46%(45%)** |
| SS (fixed-window) | −5998 | 12 002 | 0.30 | 19%(17%) | 26%(21%) |
| Random | −7430 | 14 859 | | 11%(22%) | 10%(21%) |
| *Experiment 1b* | | | | | |
| Normative | −4426 | 8833 | 0.58 | 14%(11%) | 10%(7%) |
| SS (intervention-window) | **−4054** | **8113** | 0.23 | **60%(51%)** | **58%(52%)** |
| SS (fixed-window) | −4167 | 8338 | 0.33 | 21%(17%) | 25%(25%) |
| Random | −4887 | 9774 | | 5%(21%) | 7%(17%) |
| *Experiment 2* | | | | | |
| Normative | −1058 | 2113 | 4.43 | 2%(0%) | |
| SS (intervention-window) | **−948** | **1835** | 0.19 | **50%(50%)** | |
| SS (fixed-window) | −955 | 1915 | 0.34 | 43%(40%) | |
| Random | −1059 | 2119 | | 5%(10%) | |

Note: SS refers to summary-statistic models. The "N (Regular)" and "N (Irregular)" columns display the proportion of individuals best-fit by each model according to CV, with BIC results in the brackets.

to the intervals between interventions (Table 1). This is corroborated by the individual level fits, where the largest proportion of participants were fit by *summary-statistic (intervention based)* in both regular and irregular conditions across experiments (model fits separated by conditions are shown in Table E.1).

We provide additional model fitting results in Appendix E. In Table E.2 we fit answers from Experiment 1b separated by blocks. The difference in cross-validation log-likelihood or BIC between normative and summary-statistic models was more pronounced in the no-ground-truth block than in the ground-truth block, which reflected that people's judgments were indeed more similar to the summary-statistic model. In Table E.3 we fit participants' answers with each cue separately to see whether they were dominated by Delay or Count rather than their combination. Results indicate that models with one or another cue did not fit participants' judgments better than models that mixed two cues. In Fig. E.1, we performed a grid search in $[1, 7]$ s with a step of 0.5 s to test whether the fixed-window model fits were sensitive to our choice of a 4 s window. Models with different fixed-window lengths always had substantially larger BICs than the model with the inter-intervention window approach, meaning that, even had we fit window length as an additional parameter it would not outperform by-intervention segmentation in describing participants. This was true despite the fact that the models' accuracy in causal identification is quite sensitive to the window length.

### 5.3. Discussion

In Experiment 1, we showed that people are capable of using temporal information to learn causal structures that involve generative and preventative relationships. It also showcases several interesting differences between generative and preventative causation, which we return to in the General Discussion. Human judgments were better aligned with the summary-statistic models' predictions in both quantitative results and aggregate qualitative results. Nevertheless, the data in Experiment 1 was complicated, meaning we can do more to distill simpler, more intelligible examples of how the normative and summary-statistic models diverge in their judgments. In Experiment 2, we examine judgments about minimal event sequences for which the summary-statistic and normative learners differ in their dominant answers.

## 6. Experiment 2

We designed two types of stimuli for which two models have different dominant answers. They are based on the two locality principles driving the summary-statistic model: (1) Local computation; meaning summary statistic learners fail to account for the influence of the other connections in the system, and (2) Local evidence; meaning summary statistic learners fail to take into account whatever happened before their current observation window. For the first type of stimuli we use scenarios where a learner needs to identify a generative target cause that is paired with a preventative cause. This presents a challenge for local computation because the preventative cause can block the generative causes' influence and mislead a local learner into believing the target connection is a non-causal connection, because it is statistically associated with fewer events per window or longer delays than generative causes have on average across the task. The second case type is scenarios where a non-causal target is paired with a generative neighboring component. For a local learner who only focuses on a small time window after each intervention, the generative influences can easily spill over to the observation window during which the learner is focused on the target non-causal component and leading to statistics more typical of generative causation, because it is associated with more events and shorter delays than non-causal components exhibit on average across the task. Experiment 2 focused on the regular base rate condition, since this yields the larger predicted difference between normative and summary-statistic based judgments, though we also checked that the dominant answer for each model was the same under the irregular base rate parameters.

### 6.1. Methods

#### 6.1.1. Participants

Sixty participants from Prolific were recruited and reported (32 female, 28 male, aged $41 \pm 12$). The sample size was determined by a power analysis assuming a medium sized effect ($d = 0.5$) in comparing within-subject judgments on the target cause and the goal of .90 power at the standard .05 alpha. No participants were excluded from this experiment based on the criteria we pre-registered.

#### 6.1.2. Design & procedure

Participants' task was very similar to the regular condition in Experiment 1b, where they needed to judge the roles of two connections given a 20-second clip of evidence. No video training or feedback was provided. The hand-crafted stimuli are shown in Fig. 9. For each stimulus, we call one component the "target", and the other the "lure", which could affect participants' judgments about the target. Each clip contained two segments of evidence where the two components activated close together, so their influences on the system (if any) were misleading to the summary statistic model (gray shadows in Fig. 9), but also contained evidence where the components occurred far enough apart to make the true structure recoverable by the normative model.

We constructed four exemplars of the two stimulus types (Fig. 9). For the $PG$ type (preventative lure and generative target), the lure often cancels the influence of the target, and hence the summary statistics of the target are more aligned with the non-causal summary statistics. For the $GN$ type (generative lure and non-causal target), the lure's influence spills over into the observation window of the target, leading to summary statistics more consistent with a generative target component. Therefore, the summary-statistic approach predicts systematic errors in these cases that are not predicted by the normative model (Fig. 9).

Participants went through 6 practice trials sampled from Experiment 1 (with structures $GG$, $NG$, $GP$, $NN$, $PN$, $PP$) before 8 testing trials, to ensure that they had some experience with different structures and edge types under more normal conditions. The vertical positions of two control components (above or below) were randomized across trials. The order of trials was randomized within the practice and testing phases. Participants completed 14 trials in sequence without any delineation between the practice and critical trials. The bonuses were, in reality, administered proportional to the bonuses participants gained in the practice phase (given that we predicted participants would make systematic errors in the test phase).

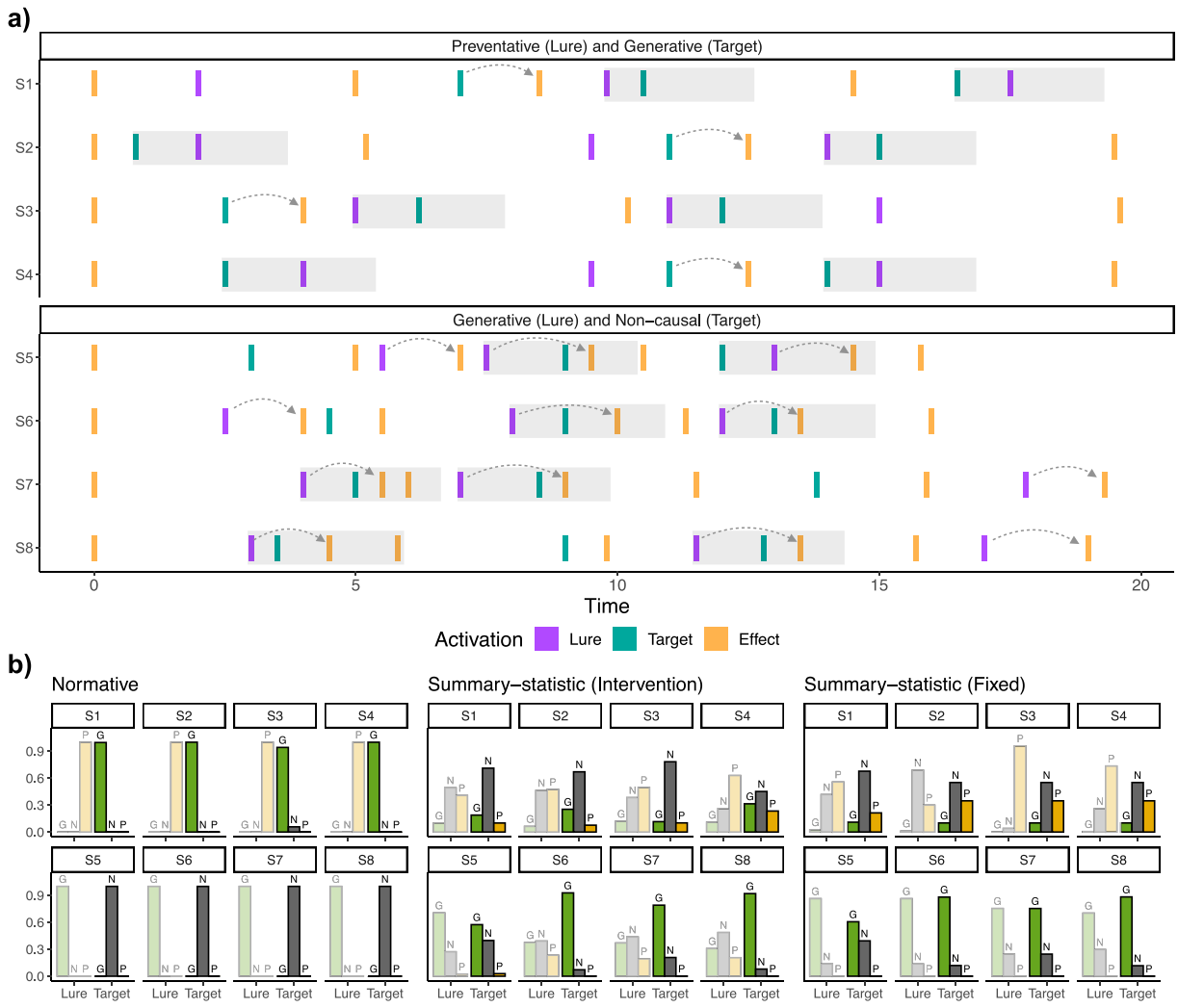


**Fig. 9.** Stimuli and model predictions in Experiment 2. (a) Stimuli. Curved arrows indicate the true underlying generative process. (b) Judgment predictions from different models. The normative and summary-statistic models particularly differ in their judgments about the target components, with opaque bars used to highlight where the modal response shifts between normative and summary statistic models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 6.2. Results & discussion

For the *PG* stimuli, participants judged the targets as non-causal $1.8 \pm 1.1$ times on average out of 4 trials (above the 33% change level, $t(59) = 3.15$, $p = .003$, $d = 0.41$, 95%CI of $d = [0.14, 0.67]$). More importantly, participants judged them more often as non-causal than generative ($t(59) = 3.62$, $p < .001$, $d = 0.82$, 95%CI of $d = [0.44, 1.19]$) or preventative ($t(59) = 2.11$, $p = .04$, $d = 0.49$, 95%CI of $d = [0.12, 0.85]$). For the *GN* stimuli, participants judged the targets as the generative one $3.1 \pm 1.0$ times on average out of 4 trials (above the 33% change level, $t(59) = 6.03$, $p < .001$, $d = 0.78$, 95%CI of $d = [0.49, 1.07]$). Meanwhile, participants judged them more often as generative than non-causal ($t(59) = 10.64$, $p < .001$, $d = 2.63$, 95%CI of $d = [2.13, 3.11]$) or preventative ($t(59) = 16.50$, $p < .001$, $d = 3.58$, 95%CI of $d = [3.00, 4.16]$). This means that for both kinds of stimuli, participants' dominant answers lined up with the summary-statistic models and diverged from those of the normative model.

The model fitting results are shown Table 1. Similar to Experiment 1, participants' answers were better fit by the summary-statistic models than the normative model. In general, they were also better aligned with the intervention-window segmentation than the fixed-window segmentation. This is also supported by a qualitative result that for *GN* stimuli, both the intervention-window model (Fig. 9) and participants (Fig. 10) regarded the lure as less likely to be a generative cause than the target component ($t(59) = 5.56$, $p < .001$, $d = 1.04$, 95%CI of $d = [0.65, 1.41]$), while the fixed-window model regarded the probabilities as more even (Fig. 9). When it comes to the individual difference, participants split more evenly across the intervention-window and fixed-window models than Experiment 1, which may imply that some participants do consider longer windows in situations when interventions interleaved heavily and hence evidence of intervention-based windows was sometimes too short to rely on.

## 7. General discussion

This paper examined how people infer causal structure on the basis of observing events in continuous time. The project was motivated by the fact that classical causal structure induction research has largely focused on inferences from atemporal statistical information, essentially sidestepping the role of event timing and delay, or else reducing it to a simple sequence of equally spaced measurements. Meanwhile, empirical research (not to mention common sense) suggests people rely strongly on event timing for causal reasoning, using temporal information to guide causal attributions even when it is inappropriate to do so.

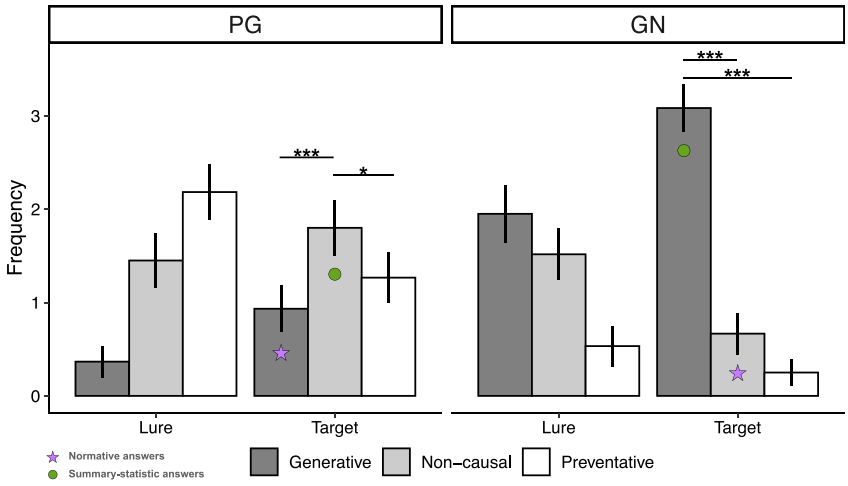


**Fig. 10.** Judgments of two types of stimuli in Experiment 2. Each type included four stimuli. Participants' dominant answers for the target component are consistent with the dominant answers from the summary-statistic model (the green dots) rather than the dominant answers from the normative model (the purple stars). Error bars indicate 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It seems likely, therefore, that time is integral to our representation of causality and hence deserves careful formal and empirical treatment.

While the space of causal structures we explored was relatively restricted, our task was challenging due to the spontaneous activations of the effect component and potential interactions between generative and preventative cause components. There were always multiple competing explanations for any effect occurrence or surprising non-occurrence, and as such, normative reasoning about the structure behind the evidence required entertaining and marginalizing over many hypothetical mappings between events. Nevertheless, participants' were able to correctly identify the majority of causal components well above chance even when base rate activations of the effect were unpredictable (Experiment 1a and 1b) and even without pretraining about the true causal delays (Experiment 1b). Our experiments thus provide an initial empirical demonstration that people can use real-time temporal information to detangle the influences of generative and preventative causes and identify causal structures involving combinations thereof.

### 7.1. Empirical findings

By including both preventative and generative relationships in our task, we have empirical results showing how the identification of these two types of relationships differ from each other in a continuous-time setting.

First, base rate regularity has a larger impact on identifying preventative relationships than generative relationships. Participants can better identify preventative connections when the effect otherwise activates regularly. This is aligned with the principle that detecting preventative causation relies heavily on one's expectation of what would otherwise have happened in the causal system (Buehner et al., 2003; Cheng, 1997; Griffiths & Tenenbaum, 2005).

Second, when judging a causal connection in the system, the type of neighboring connections matters. Experiment 1 showed that when the base rate is irregular, participants could better identify a connection when it was paired with a generative neighbor rather than a non-causal or preventative neighbor. This can be explained by the fact that a generative connection can increase the predictability of the effect, which is helpful in general but particularly when the base rate is unpredictable. Experiment 2 showed that a preventative neighbor can cancel out a generative influence and mislead people to judge a generative connection as non-causal.

Third, the timing and sequence of interventions matter when making causal judgments, and it affects the identification of generative and preventative connections in different ways. Participants identified generative and non-causal relationships better when the interventions were clustered, rather than interleaved. This makes sense given that the evidence under clustered interventions involves less interference from neighboring connections. We confirmed this in Experiment 2 where we show that deliberately interleaved evidence leads participants to systematically mistake the roles of generative and non-causal connections. In contrast, the advantage of clustered interventions disappeared when it came to prevention. To identify preventative relationships, it makes sense to spread out interventions so their influence covers more of the timeline, and in particular to perform them ahead of whenever one has a strong expectation of the effect occurring (Lovibond & Lee, 2021; Melchers et al., 2006). To our knowledge, these findings represent the first systematic investigation of how human causal judgments engage with a setting where generative and preventative causal influences intertwine and interact in time.

### 7.2. Normative vs. summary-statistics

To better understand how participants made their judgments, we contrasted two learning models: An exhaustive normative account and a summary-statistic-based local approximation. Both accounts were able to identify generative and preventative influences well in our task, but only the summary statistic account could capture cases in which participants were worse at identifying the non-causal connections (Experiment 1) and misled by interleaved interventions (Experiment 1 and 2). Quantitatively, the summary-statistic account also fits participants' judgments across both experiments better.

Our normative model demonstrates that near perfect inversion of the generative causal model is possible for a learner with exactly the correct delay assumptions and unlimited processing power. It works via reasoning at the token level of actual attribution (Halpern, 2016), suggesting this kind of reasoning is key for achieving benchmark performance in this small data setting. The summary-statistics account takes a different approach that is computationally much more frugal and scalable to more complex causal models, but has the cost of being less sensitive to precise event timing, and being more susceptible to interference between components. The approach combines several core principles of bounded cognitive processing: Use of simulation from generative mental models and comparison via summary statistics in place of an exact or intractable likelihood calculation (Battaglia et al., 2013; Blum et al., 2013; Lintusaari et al., 2017; Sunnåker et al., 2013;

Ullman et al., 2018). It combines this with local (Bramley, Dayan et al., 2017; Davis et al., 2020; Fernbach & Sloman, 2009) and incremental (Bramley, Dayan et al., 2017; Davis et al., 2020; Rehder et al., 2022) processing to break up the global inference problem into a series of spatially and temporally local subproblems. The departures from the ideal of the global normative thinker allow it to explain several error patterns exhibited by participants. In general the normative model serves to showcase the rapidly compounding challenge of maintaining a global perspective when processing evidence that includes multiple causal influences that intertwine and interact in real time (Bramley, Mayrhofer, Gerstenberg and Lagnado, 2017; Gong et al., 2023).

Imagined experiences are a core feature of our conscious experience and as such, mental simulation has been implicated by a number of theories of cognition as playing key roles in both model-based inference and planning (Battaglia et al., 2013; Bramley, Gerstenberg, Tenenbaum and Gureckis, 2018; Gerstenberg et al., 2021; Hamrick et al., 2016; Ludwin-Peery, Bramley, Davis, & Gureckis, 2020; Ullman et al., 2018). Mental simulation is thought to be key to offline (Hinton, Dayan, Frey, & Neal, 1995), and simulation phases are now a common part of the training regimen for large reinforcement learning models (Ellis et al., 2020; Mnih et al., 2015). Our experiments add one small piece to this research line, showing how an inference mechanism grounded in simulation and the extraction of summary statistics may explain how people mitigate the computational costs involved in reverse engineering the causal mechanisms that explain the events we observe in real time.

The idea of combining sampling from a generative model with summary statistics stems from Approximate Bayesian Computation (Blum et al., 2013; Lintusaari et al., 2017; Sunnåker et al., 2013). The approach makes it possible to approximate an intractable Bayesian inference by using the similarity between data simulated from a hypothesized model or parameter setting and observed data as a proxy for the likelihood of that model or parameter setting. Choosing the best summary statistics or loss function for a domain is a research area in itself in machine learning (Csilléry, Blum, Gaggiotti, & François, 2010), while identifying what summary statistics might be used in cognition is another challenging and unsolved problem. We do not solve this problem here, but simply hand selected two basic summary statistics (cf. Ullman et al., 2018) on the grounds that they reflect the most basic and easily reported timing measurements people can make in online settings. We showed that the delay and count cues were reasonably diagnostic in our task (Experiment 1) but also unpacked the circumstances under which they can be misleading (Experiment 2).

Within the summary-statistic framework, we considered two ways participants might segment the trials into counting windows. We proposed they might either track events within fixed-length windows after each intervention or use the gaps between each intervention directly as a count window. The inter-intervention segmentation variant captured participants' behavior better despite the fact that the windows were of markedly different lengths detracting from the reliability of the metric. A potential explanation for this is that people may be fundamentally unable to track events from multiple causal perspectives in parallel, thus being forced to rely on the uneven inter-event windows (Bramley, Gerstenberg, Mayrhofer et al., 2018; Davis et al., 2020). Of course, in an active learning context, the learner is free to perform interventions at their own pace. This research suggests that what learners are able to attend to and measure is likely to shape their approach to interventions in time. For instance, one way to make inter-intervention count statistics as powerful as possible is to intervene on a regular schedule, eliminating the confound of episode length, while leaving as large as possible gaps between interventions additionally minimizes spillover effects. Interestingly, these are cognitive rather than normative considerations since the ideal observer is practically indifferent to the regularity of the intervention spacing.

### 7.3. Alternative accounts

One popular recent idea in the causal cognition literature is that people form and adjust causal theories locally and incrementally (Bramley, Dayan et al., 2017; Bramley, Mayrhofer et al., 2017; Davis et al., 2020; Fernbach & Sloman, 2009; Markant et al., 2016). For instance, Bramley, Dayan et al. (2017) model causal structure learning (in discrete trial contexts) as a process of incremental adaptation of a single global hypothesis driven by the need to accommodate new evidence as it arrives. They argue that causal learners do focus locally when grappling with complex structures, but that many are able to condition on their current beliefs about neighboring connections rather than ignoring them altogether, leading to patterns of sequential local focus and anchoring that still tend to favor the correct global structure in the limit. We did not collect the interim judgments we would need to probe this account directly, but we think it is entirely plausible that people focused on the roles of the components not just separately but also serially, perhaps flipping their attention back and forth several times throughout a trial. For example, if participants focused on a generative component first and a preventative component second, they might have been able to take advantage of their expectation of events produced by the apparently generative component to supercharge their inferences about prevention.

The other idea is based on the "smart initialization and short search" algorithm in Ullman et al. (2018). Analogous to our findings, they showed that although human physical learning was better captured by a summary-statistic account than a noisily normative Bayesian model, responses could be even better fit by a mechanism that combines the two. Their best-fit model used the prediction of a summary-statistic approach as a starting hypothesis, and then made local adjustments to this by running a short Markov Chain Monte Carlo search chain. Such a smart initialization could play an important role here too. It is plausible that some participants may have performed similar steps, i.e. forming an impression of the role of a component due to the delays and counts but adjusting this when accommodating a belief about the neighboring connection or an understanding of the regularity of the base rate.

### 7.4. Future directions

To date, causal learning in continuous time has received little attention, meaning there are numerous basic research questions still to be addressed. In the current paper, we focus on just one of these, providing a close examination of the interplay between inference about generative and preventative causal relationships. However, for this we make specific assumptions about the scope with which preventative influences work. Concretely, we conceive of preventative influences as eliminating all expected effects for a short time no matter their cause. However, there are several alternatives that seem at least as salient and may be more appropriate depending on knowledge of the context and mechanisms involved. For example, prevention could work by blocking the next one event (or perhaps the next $N$ events) rather than blocking everything for a fixed window. Prevention could also operate on "links" rather than "nodes" within the causal graph, for example blocking the action of a generative cause on an effect, but leaving the spontaneous activations of that effect intact, or visa versa (Carroll & Cheng, 2009; Chow, Lee, & Lovibond, 2023; Fraser & Holland, 2019).

In the current learning task, causal influences were represented as operating between point events. This is a major simplification from many real scenarios in which variables involved in causal interactions are often able to take multiple, or even a continuum of, values. The cat in our motivating example might drink more or less water or hold different teaser toys in higher or lower regard leading to faster, slower, more or less intense effects. Even though events are abstractions of continuous inputs, and many, such as state changes, are readily thought of as punctate, many everyday event concepts clearly have non-zero duration and often have internal structure such as a gradual

or sudden onset or offset. For example, given enough time, many of the states referred to in causal learning scenarios are not permanent. "Wet ground" dries. "Tanned skin" fades. Many disequilibria will either dissipate or recover without external intervention. Other states, such as a turned-on light bulb may tend to persist until canceled, i.e. by switching the switch a second time. These could be seen as events with an infinitely long duration (i.e. permanent state-changes). As event duration reduces, it becomes less likely that events will overshadow one another. Point events are a limiting-case abstraction of this where the duration is reduced to zero, resulting in a setting where there is no true causal overshadowing (Paul & Hall, 2013). That is, generative cause will always produce an observable effect even if it occurs close to another event. However, in settings with longer events it becomes increasingly important to consider the super-secession situations and perhaps to apply the noisy-or or noisy-and-not frameworks (Cheng, 1997; Griffiths & Tenenbaum, 2009) that capture how in contingency settings, effects can easily be hidden due to an already-occurring, or already-prevented target. Future research could study how people represent the duration of causal events as well as their influences and thus begin to form a richer theory of causal concepts in time that captures a wider range of relata, variables, influences, and events.

Finally, we focused on online causal learning here, where information flowed in rapidly and learners had no opportunity to replay and revise. However, it is possible that people are capable of reasoning more normatively in offline learning tasks when they are provided with information summarized in a timeline and can take as long as they like to consider the fit between the data and different causal hypotheses (Bramley, Gerstenberg, Mayrhofer et al., 2018). Furthermore, to the extent that summary-statistic based inference and normative inference deviate, it seems likely that people's judgments after additional thinking time could differ from their more instinctive or gut responses (Ludwin-Peery et al., 2020). Reflective thinking has been studied for decades in human reasoning and decision making (Kahneman, 2011; Sloman, 1996), while it is less studied in causal inference. The normative vs. summary-statistic contrast in this paper provides a potential paradigm for operationalizing the role of reflective thinking in causal inference.

### 7.5. Conclusions

In this paper, we showed that people can use information in continuous real time to learn about causal systems that potentially contain generative and preventative causal relationships. Their performance was influenced by multiple factors, including the nature of the causal influences (generative, non-causal, preventative), interactions with neighboring connections, base rate regularity, and intervention patterns. We laid out both a normative framework and a process-level model. Both qualitatively and quantitatively, human judgments were better captured by the process-level summary-statistic account, capturing the idea that people may infer causal structure via statistical cues such as average delays and counts that are much easier to track in real time than the exact generative model likelihoods. This work thus provides a quantitative account of how people manage to learn causal structure, in particular preventative influences, on the basis of continuous temporal dynamics. This contributes to our understanding of natural cognition and sheds light on the challenging question of how any cognitive agent can succeed in forming an internal causal model of a complex and continuous environment.

## CRediT authorship contribution statement

**Tianwei Gong:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft. **Neil R. Bramley:** Conceptualization, Methodology, Formal analysis, Supervision, Funding acquisition, Writing – review & editing.

## Data availability

The pre-registrations, materials, data, and analysis code are available at https://osf.io/q8n72/. Stimuli for all experiments can be viewed at https://github.com/tianweigong/causal_diamond.

## Appendix A. Gamma distributions

We use the gamma distributions to generate causal delays for experimental stimuli, and also to model causal inferences over time (Bramley, Gerstenberg, Mayrhofer et al., 2018; Bramley, Mayrhofer et al., 2017; Stephan et al., 2020; Valentin et al., 2022). Gamma distributions Gamma$(\alpha, \beta)$ define a density over $(0, +\infty)$ with two parameters — shape $\alpha$ and rate $\beta$— controlling the expectation and central tendency of the delay (mean $\mu = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$, see Fig. 2 for examples). Memoryless exponential distributions are special cases of gamma distributions when $\alpha = 1$, where the expected delay is constant, no matter how long you have already waited for (Fig. 2d). This is useful for representing occurrence of events generated by unknown and unobserved background causes. Gamma distributions also have a convenient transition property that facilitates the calculation of preventative causation: if $X, Y \sim \text{Gamma}(\alpha, \beta)$ then $X + Y \sim \text{Gamma}(2\alpha, \beta)$. As an example, suppose a bus arrives every $12 \pm 2$ min. If you arrive at the bus stop just as it leaves you might expect to wait $\text{Gamma}(\alpha : 36, \beta : 3, [\mu : 12, \sigma : 2])$ minutes. However, if you then are told the next bus is canceled, the expected waiting time will double while the variance will increase rather less: $\text{Gamma}(\alpha : 36, \beta : 3) + \text{Gamma}(\alpha : 36, \beta : 3)$ is equal to $\text{Gamma}(\alpha : 72, \beta : 3, [\mu : 24, \sigma : 2.8])$. We will take advantage of this feature in building our normative inference model.

## Appendix B. Normative calculations

The normative learner updates the prior over structures $P(S)$ (here assumed to be uniform), with a likelihood function to obtain a posterior distribution, given the set of gamma parameters $\mathbf{w}$ which indicates the belief about delays:

$$P(S|\mathbf{d}, \mathbf{w}; \mathbf{i}) \propto p(\mathbf{d}|S, \mathbf{w}; \mathbf{i}) \cdot P(S) \tag{1}$$

Here $\mathbf{d}$ refers to effect data (E's activations), which is conditioned upon a set of interventions $\mathbf{i}$ on the causes (A or B).

In order to maintain rational beliefs about causal structure, the ideal reasoner considers all possible causal paths $\mathbf{Z}_s$ that could describe what actually happened given each possible structural hypothesis $s \in \mathbf{S}$, summing up the individual likelihood of these mutually exclusive and exhaustive possibilities to assess the overall likelihood of each structure hypothesis:

$$P(\mathbf{d}|s, \mathbf{w}; \mathbf{i}) = \sum_{z' \in \mathbf{Z}_s} P(z'|s, \mathbf{w}; \mathbf{i}) \tag{2}$$

Normative causal attribution involves three steps: (1) attributing causes to effects that have occurred; (2) explaining away effects that should or might have occurred but were not observed; (3) examining the temporal distance between presumed preventative events and the subsequent effect event. Step 1 and 2 correspond to path construction. We use $\{\alpha_g, \beta_g\}$, $\{\alpha_p, \beta_p\}$, $\{\alpha_b, \beta_b\}$ to denote parameters of gamma distributions for generative delays, preventative windows, and base rate delays. In the current experiments: $\{\alpha_g = 9, \beta_g = 6\}$, $\{\alpha_p = 36, \beta_p = 12\}$, and $\{\alpha_b = 100, \beta_b = 20\}$ (regular base rate) or $\{\alpha_b = 1, \beta_b = 0.2\}$ (irregular base rate).

Step 1 is to form $g' \rightarrow e'$ pairs where (1) the effect event $e'$ is not over-determined (i.e. has a single actual cause), (2) the cause event $g'$ does not produce its effect twice, and (3) $g'$ precedes $e'$. The likelihood of each pair is then determined by mapping the delay between $g'$ and $e'$ to the gamma density function:

$$P(g' \rightarrow e'|\alpha_g, \beta_g) = P(t_{g' \rightarrow e'} = t_{g'e'}|\alpha_g, \beta_g) \tag{3}$$

Step 2 involves forming $g' \rightarrow h$ pairs where $h$ is a hidden effect event assumed to happen sometime after the observable period *or* at some point during a preventative window. The likelihood calculation depends on the gamma cumulative density falling beyond the end of the clip or within the window:

$$P(g' \rightarrow h|\alpha_g, \beta_g, \alpha_p, \beta_p) = P(t_{g'\rightarrow h} > t_{end}|\alpha_g, \beta_g) + P(t_{g'\rightarrow h} \leqslant t_{end}|\alpha_g, \beta_g)(1 - \prod_{p'}(1 - P(t_{g'\rightarrow h} < t_{g'} + t_{p'\rightarrow h}|\alpha_g, \beta_g, \alpha_p, \beta_p))) \quad (4)$$

Base rate activations of the effect event are represented as having been caused by the previous base rate activation, which can also be represented as $g' \rightarrow e'$ pairs where $g'$ is actually the target component's (i.e. E) activation. When there are presumed preventative cause events, the base rate activation could be prevented but then subsequently "recover". Therefore, for base rate activation we could jointly consider Step 1 and Step 2 as $g' \rightarrow h^{(1)} \rightarrow \cdots \rightarrow h^{(n)} \rightarrow e'$, where $h^{(1)} \ldots h^{(n)}$ happens within the preventative windows. Meanwhile, according to the transition property of the gamma distribution (see Appendix A), if $X, Y \sim Gamma(\alpha, \beta)$ then $X + Y \sim Gamma(2\alpha, \beta)$. The probability $P(g' \rightarrow h^{(1)} \rightarrow \cdots \rightarrow h^{(n)} \rightarrow e')$ can thus be represented as Eq. (5), where the calculation of $P(g' \rightarrow e')$ is similar to Eq. (3), and the calculation of $P(g' \rightarrow h^{(n')})$ is similar to Eq. (4) except that $t_{end}$ is substituted with $t_{e'}$ and only the second item of prevention is considered.

$$P(g' \rightarrow h^{(1)} \rightarrow \cdots \rightarrow h^{(n)} \rightarrow e'|\alpha_b, \beta_b, \alpha_p, \beta_p) = P(g' \rightarrow e'|(n+1)\alpha_b, \beta_b) \prod_{n' \in n} P(g' \rightarrow h^{(n')}|n\alpha_b, \beta_b, \alpha_p, \beta_p) \quad (5)$$

Finally, the prevention examination in Step 3 extracts all presumed preventative events and their nearest effect events to form $p' \rightarrow e'$ pairs (there is no need for examination if no effect events happen after $p'$), and then applies the gamma cumulative density function of prevention:

$$P(p' \rightarrow e'|\alpha_p, \beta_p) = P(t_{p'\rightarrow e'} < t_{p'e'}|\alpha_p, \beta_p) \quad (6)$$

## Appendix C. Implementation of simulation-and-summary-statistic models

### C.1. Cue distributions

We constructed the cue distributions (see Fig. 3b) for each type of connection (generative, non-causal, preventative) under two base rates (regular, irregular) by simulating 90,000 interactions with imagined causal devices. These included 10,000 simulations of each of the 9 causal structures considered here. In each simulation the structure is perturbed by interventions performed in random orders with random timings.[6] In this way we establish a marginal distribution for each summary statistic under each type of connection. Note that we used a large number of simulations to produce smooth distributions for our later model fitting, however similar distributions can be achieved with a much smaller number of simulations (Ullman et al., 2018). As shown in Fig. 3b, the delay cue is independent of questions of segmentation by definition since it always relates to the earliest subsequent effect event after each intervention. The count cue, however, is sensitive to the choice of segmentation, meaning we consider intervention-window and fixed-window assumptions separately. For delay distributions, we use a probability density function smoothed with Gaussian kernels, while for count distributions we can use the discrete probability mass functions directly.

[6] Similar to generating the experimental stimuli, each simulation included three interventions on A and three interventions on B. Distinct from the experimental stimuli, simulated sequences here were not cut at twenty seconds so as to avoid the complex boundary effects in distribution construction.

### C.2. Likelihood calculation

We assume each connection is estimated independently as either generative, non-causal, or preventative, and then combined to yield an overall probability for each candidate causal structure. For example, an intervention on $A$ with the nearest effect occurring 2.5 s later has a likelihood of [.2, .7, .1] of having been produced by a generative, non-causal or preventative $A \rightarrow E$ connection respectively under the regular base rate and [.3, .6, .2] under the irregular base rate. When the next intervention on $A$ happens, the posterior is updated by taking the product of this new likelihood with the preceding ones.

### C.3. Boundary situations

We consider boundary situations when observing evidence as follows: If no effect occurs within the observation window, in both segmentation approaches, the delay cue will be marked as larger than the observation window and the probability is estimated according to the cumulative density function falling after this. If the observation window is less than the fixed window length for the fixed-window approach (which often happens near the end of the clip), or there is no next intervention in the intervention-window approach, the count cue will be marked as greater than or equal to the observed count of effects and the probability is also estimated on the basis of its cumulative mass function.

## Appendix D. Model fitting procedure

We considered four models in total:

1. Fully normative inference based on marginalizing over all possible causal pathways.
2. Summary-statistic (SS) based inference, using a fixed 4 s window to count events following each intervention.
3. Summary-statistic based inference, using the interval until the next intervention to count events.
4. A parameter free baseline that predicts all structure judgments to be selected with equal probability.

As in our comparison to simulations, we simply assume the delay and count cues are equally weighted and merged. We assume learners begin each problem with a uniform prior over causal structures. We feel this is a reasonable choice here since the relatively small hypothesis space,a balanced set of trials, and the abstract setting leave little for inductive biases to attach to. Nevertheless, we accept that we cannot rule out the possibility that some of the findings we attribute to evidence processing enter through prior preferences. To map models' posterior probabilities to judgments, we assumed participants' responses result from a softmax over a posterior probability vector $v$:

$$P(n) = \frac{\exp(v_n/\tau)}{\sum_{n' \in N} \exp(v_{n'}/\tau)} \quad (7)$$

The "temperature" parameter $\tau \in (0, +\infty]$ controls how reliably the participant selects the most probable answer (i.e. that with the largest $v_n$ in choice $n$). Smaller $\tau$ connotes higher choice reliability with $\tau = 0$ corresponding to hard maximization and $\tau \rightarrow \infty$ approaching random responding.

We evaluate model fit using cross-validation. At the aggregate level, we fit parameters to the judgments from $K - 1$ subsets of the complete dataset, and evaluate model performance in terms of its log-likelihood of predicting the left-out subset. $K$ was defined via the stimulus seeds in each experiment (i.e. $K = 18$ in Experiment 1a and $K = 12$ in Experiment 1b including stimuli with and without a ground truth). This provides a rigorous and generalizable test of the models, since the actual sampled values of the stimuli (e.g. intervention timing, base rate activating timing, etc.) are always outside of the training sample

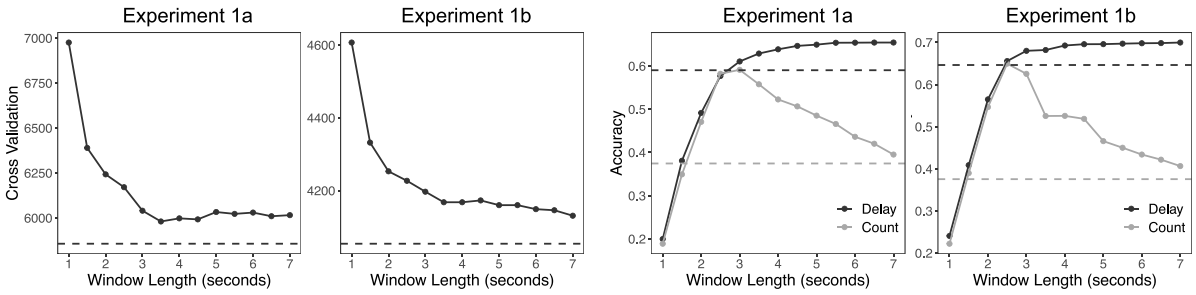


**Fig. E.1.** Cross validation results and model accuracy under different fixed-window lengths for summary-statistic models. Horizontal dashed lines indicate cases of intervention-window segmentation.

**Table E.1**
Model fits separated by conditions.

| | Regular | | | Irregular | | |
|---|---|---|---|---|---|---|
| | CV | BIC | $\tau$ | CV | BIC | $\tau$ |
| *Experiment 1a* | | | | | | |
| Normative | −2894 | 5789 | 0.45 | −3162 | 6327 | 0.43 |
| SS (intervention-window) | **−2822** | **5648** | 0.23 | **−3036** | **6076** | 0.22 |
| SS (fixed-window) | −2924 | 5853 | 0.31 | −3074 | 6153 | 0.29 |
| Random | −3695 | 7390 | | −3735 | 7469 | |
| *Experiment 1b* | | | | | | |
| Normative | −2256 | 4497 | 0.62 | −2167 | 4332 | 0.51 |
| SS (intervention-window) | **−2041** | **4086** | 0.23 | **−2014** | **4032** | 0.24 |
| SS (fixed-window) | −2114 | 4232 | 0.35 | −2052 | 4106 | 0.31 |
| Random | −2503 | 5006 | | −2384 | 4768 | |

**Table E.2**
Model fits separated by blocks in Experiment 1b.

| | Ground truth | | | No ground truth | | |
|---|---|---|---|---|---|---|
| | CV | BIC | $\tau$ | CV | BIC | $\tau$ |
| Normative | −2009 | 4022 | 0.46 | −2361 | 4725 | 0.88 |
| SS (intervention-window) | **−1917** | **3840** | 0.24 | **−2141** | **4279** | 0.23 |
| SS (fixed-window) | −1982 | 3969 | 0.32 | −2188 | 4373 | 0.35 |
| Random | −2443 | 4887 | | −2443 | 4887 | |

**Table E.3**
Model fits with one cue.

| | Delay | | | Count | | |
|---|---|---|---|---|---|---|
| | CV | BIC | $\tau$ | CV | BIC | $\tau$ |
| *Experiment 1a* | | | | | | |
| SS (intervention-window) | −5994 | 11 990 | 0.31 | −6065 | 12 134 | 0.20 |
| SS (fixed-window) | −6040 | 12 084 | 0.35 | −6228 | 12 460 | 0.33 |
| *Experiment 1b* | | | | | | |
| SS (intervention-window) | −4136 | 8277 | 0.33 | −4173 | 8343 | 0.20 |
| SS (fixed-window) | −4196 | 8393 | 0.39 | −4292 | 8585 | 0.36 |

for all test sets. On the individual level, we similarly applied hold-one-stimulus-out as our cross-validation scheme for all experiments. For easy familiarity and comparability with other model based analyses of causal learning data, we also report Bayesian Information Criterion (BIC) penalized fits to the full dataset.

## Appendix E. Alternative model fitting results

See Tables E.1–E.3 and Fig. E.1.

## References


Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Psychology Press.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 238–249.

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology*, *53*(3), 155–167.

Blum, M. G., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*(2), 189–208.

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

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning.. *Psychological Review*, *124*(3), 301–338.

Bramley, N. R., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 236–241).

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

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

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

Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 150–155).

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119–1140.

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

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

Carroll, C., & Cheng, P. (2009). Preventative scope in causation. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the cognitive science society* (pp. 833–838).

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

Chow, J. Y., Lee, J. C., & Lovibond, P. F. (2023). Inhibitory learning with bidirectional outcomes: Prevention learning or causal learning in the opposite direction? *Journal of the Cognition*, *6*(1), 1–24.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418.

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

Davis, Z., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), Article e12839.

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., et al. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. arXiv preprint arXiv:2006.08381.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 678–693.

Fraser, K. M., & Holland, P. C. (2019). Occasion setting. *Behavioral Neuroscience*, *133*(2), 145–175.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936—975.
Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744.
Gong, T., & Bramley, N. R. (2020). What you didn't see: Prevention and generation in continuous time causal induction. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42th annual conference of the cognitive science society* (pp. 2908–2914).
Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, Article 101542.
Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119.
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*(5), 620–629.
Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.
Grice, G. R. (1948). The relation of secondary reinforcement to delayed reward in visual discrimination learning. *Journal of Experimental Psychology*, *38*(1), 1–16.
Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.
Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.
Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.
Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*(7), 1128–1137.
Halpern, J. Y. (2016). *Actual causation*. MIT Press.
Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.
Harman, G. (1986). *Change in view: principles of reasoning*. The MIT Press.
Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.
Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*(5214), 1158–1161.
Hume, D. (1740). *A treatise of human nature oxford*. New York: Oxford University Press (2000 reprint).
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.
Lagnado, D. A. (2011). Causal thinking. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 129–149). New York: Oxford University Press.
Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876.
Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.
Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*(1), 184–195.
Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: psychology, philosophy, and computation* (pp. 154–172). New York: Oxford University Press.
Lee, J. C., & Lovibond, P. F. (2021). Individual differences in causal structures inferred during feature negative learning. *Quarterly Journal of Experimental Psychology*, *74*(1), 150–165.
Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, *66*(1), e66–e82.
Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.
Lovibond, P. F., & Lee, J. C. (2021). Inhibitory causal structures in serial and simultaneous feature negative learning. *Quarterly Journal of Experimental Psychology*, *74*(12), 2165–2181.
Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.
Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge: MIT Press.
Melchers, K. G., Wolff, S., & Lachnit, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review*, *13*(4), 662–667.
Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology*, *22*(3), 408–412.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.
Pacer, M., & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 833–838).
Pacer, M., & 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 conference of the cognitive science society* (pp. 1805–1810).
Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.
Pearl, J. (2000). *Causality*. New York: Cambridge University Press (2009 reprint).
Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York: Basic Books.
Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, and correlation. *Journal of the Machine Learning Technologies*, *2*(1), 37–63.
Rehder, B., Davis, Z. J., & Bramley, N. (2022). The paradox of time in dynamic causal systems. *Entropy*, *24*(7), 863.
Rescorla, R. A., & Wagner, A. R. (1972). A theory on pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. Black, & W. Prokasy (Eds.), *Classical conditioning ii: current theory and research* (pp. 64–99). New York: Appleton Century Crofts.
Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1233–1256.
Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139.
Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1–2), 93–125.
Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1–21.
Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, *19*(4), 353–360.
Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139–159.
Simon, H. A. (1982). *Models of bounded rationality: empirically grounded economic reason*. MIT Press.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.
Sloman, S. A. (2005). *Causal models: how people think about the world and its alternatives*. Oxford University Press.
Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation–A computational model. *Cognitive Science*, *44*(7), Article e12871.
Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Computational Biology*, *9*(1), Article e1002803.
Tarpy, R. M., & Sawabini, F. L. (1974). Reinforcement delay: A selective review of the last decade. *Psychological Bulletin*, *81*(12), 984–997.
Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.
Valentin, S., Bramley, N. R., & Lucas, C. G. (2022). Discovering common hidden causes in sequences of events. *Computational Brain & Behavior*, 1–23.
Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, *25*(A), 287–298.
Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*(2), 92–97.
Zhao, Y., Zeng, T., Wang, T., Fang, F., Pan, Y., & Jia, J. (2023). Subcortical encoding of summary statistics in humans. *Cognition*, *234*, Article 105384.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223.