Learning preventative and generative causal structures from point events in continuous time

Tianwei Gong
Department of Psychology
University of Edinburgh
Edinburgh, Scotland EH8 9JZ
tia.gong@ed.ac.uk

Neil R. Bramley
Department of Psychology
University of Edinburgh
Edinburgh, Scotland EH8 9JZ
neil.bramley@ed.ac.uk

Abstract

Many previous accounts of causal structure induction have focused on atemporal contingency data while fewer have described learning on the basis of observations of events unfolding over time. How do people use temporal information to infer causal structures? Here we develop a computational-level framework and propose several algorithmic-level approximations to explain how people impute causal structures from continuous-time event sequences. We compare both normative and process accounts to participant behavior across two experiments. We consider structures combining both generative and preventative causal relationships in the presence of either regular or irregular background noise in the form of spontaneous activations. We find that 1) humans are robustly capable learners in this setting, successfully identifying a variety of ground truth structures but 2) diverging from our computational-level account in ways we can explain with a more tractable *simulation and summary statistics* approximation scheme. We thus argue that human structure induction from temporal information relies on comparisons between observed patterns and expectations established via mental simulation.

1 Introduction

We naturally think about the world in terms of the progression of events that cause and affect one another. Causal reasoning helps us abstract from a real-time experience to stable causal mechanisms that we can use to explain, predict and sometimes control our environment. How do people form these causally structured representations? What algorithms best capture human-like causal learning from evidence arriving in continuous time? Research has focused on several aspects of human causal learning, including 1) learning the form of generative and preventative relationships [1], 2) distinguishing relationships from spurious correlations [2]; 3) inferring causal structure across multiple relata [3]; 4) leveraging temporal order and delay information [4]. However, while all these dimensions are inseparable in real-world causal learning [5], to our knowledge no study has yet combined them in a single learning task. Previous studies of temporal structure induction have typically focused on generative relationships, while quantitative accounts of both generative and preventive causation have been restricted to contingency settings and frequently to pairwise relationships.

In this paper, we propose a causal learning task in which causal systems unfold over time due to potentially generative and preventative causal relations against a background of exogenous influences. We formulate a normative model of exact inference in this setting but note that it scales poorly, becoming computationally infeasible when reasoning about more than a handful of events. Thus we posit that in practice, human learners adopt more efficient approximations [6]. As such, we consider a class of process-level algorithms those based on establishing easy-to-measure features of

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

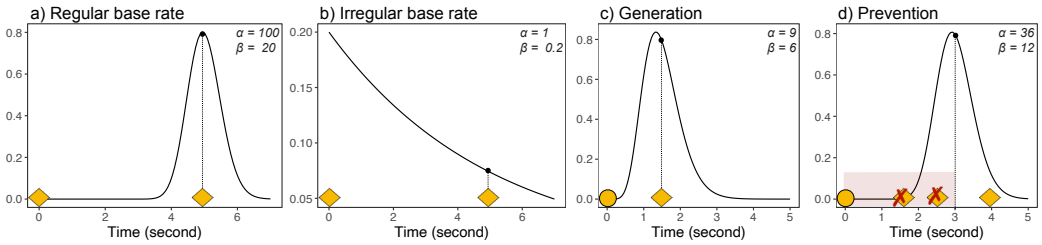

Figure 1: Gamma density distributions. (a) A regularly spontaneously occurring event. (b) A randomly spontaneously occurring event. (c) The effect of a generative cause and (d) the effect following a preventative cause. Circles indicate cause events and diamonds indicate effect events.The shaded area indicates the prevention window during which any activation of the effect would be blocked.

generative, preventative and non-causal relationships and using these as indirect cues to structure [7]. We compare this approach against exact inference, finding it provides the more compelling account of human responses across two experiments.

2 Background

Prevention and generation in causal structure learning There have been numerous studies and models describing causal learning from contingency information — i.e., patterns of co-occurrence of discrete variable states across independent trials. Early accounts of causal cognition focused modeling judgments of the strength of a causal relationship between pairs of (usually binary) variables. This research distinguished between two basic forms of causal power — generative power, whereby the presence of a cause increases the probability of its effect occurring and preventative, whereby the presence of the cause decreases the probability of its effect. ΔP [8] — meaning the change in the probability of an effect occurrence with vs. without the cause ($\Delta P = P(E|C) - P(E|\neg C)$). It provides a basic index for the strength and direction of such effects. However, it turned out to be a poor model of human causal judgments from contingencies. People are additionally sensitive to the base rate of the effect $P(E|\neg C)$, i.e., how frequently the effect occurs in the absence of the cause. Cheng's Power PC theory captured this sensitivity [1] by assuming people intuitively partial out the influence of background causes when making judgments of the strength of a focal causal relationship. This conceptual move explains findings that under fixed ΔP, people infer stronger generative influences when base rates are high, while inferring stronger preventative influences when the base rates are low [9]. In extreme cases, people regard evidence as uninformative with respect to generative judgments if the base rate is 1 — the effect always occurs anyway, so there are no opportunities to see if the cause was effective – and with respect to preventative judgments if the base rate is 0 — where there is nothing for the causal influence to prevent [10].

Incorporating reasoning about background causes implies that local causal judgments are not made in isolation, but rather take surrounding structure into account. Indeed, Griffiths and Tenenbaum [2] show that human causal judgments from contingencies in experiments reflect rational inferences about the probability that a particular causal relationship exists at all, as well as its functional form, essentially distinguishing between candidate causal Bayesian network structures, one with a connection from C to E and one without. This combination of directed acyclic graphs (DAGs) with Bayesian inference also allows rational models of structure learning for arbitrary numbers of relata. The framework of causal Bayesian networks [11] further enables incorporation of other forms of evidence such as that resulting from interventions on the system of interest and use of the resulting model for hypothetical and counterfactual thinking. By using these tools, researchers have built models to explain human judgments and active learning strategies in contingency-based structure learning [3, 12, 13], albeit focusing predominantly on generative relationships.

Time in causality Besides contingency information, time also plays an important role in human causal reasoning. Bringing temporal information into causal learning thus adds to the ecological validity of experiments, since causes often take time to reveal their effects in reality [14]. Generally, people make stronger causal attributions for short temporal delays than long temporal delays [15], but

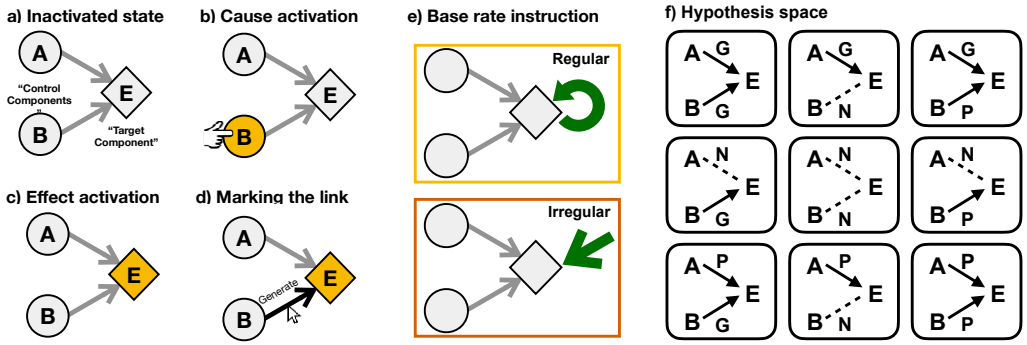

Figure 2: Causal devices tested in this paper. a-d) Experimental interfaces. e) the setting of regular vs. irregular base rate. f) The response hypothesis space (all possible causal structures where G = generative; N = non-causal; P = preventative).

this is moderated by expectation, with shorter-than-expected delays also reducing causal judgment strength [16]. People are also sensitive to delay reliability with causal judgments decreasing as increasing interval variability between putative causes and effects increases [4, 17]. In order to capture a general preference for the predictable and reliable delays, we can model causal influences over time using gamma distributions. Gamma distributions Gamma(α, β) define a density over $(0, +\infty)$ with two parameters controlling the expectation and central tendency of the delay (e.g. with a shape α and rate β, see Figure 1). Memoryless exponential distributions are special cases of gamma distributions when $\alpha = 1$, where the expected delay remains constant, no matter how long you have already waited for (Figure 1b).

Several recent studies have adopted gamma distributions directly (or indirectly via the mathematically-related Poisson process) to model temporal causal representation and reasoning, including between pairs of events [18], for structure learning [4], imputing hidden causes [19], and making judgments of actual causation given a known causal structure [20]. However, as with contingency work, these studies have largely focused on cases of generative causal influence. Additionally, past studies have focused on inference from sets of independent clips, in which root components are usually activated at the start and effects followed. Yet a more realistic setting is where causes and effects to intermingle and components can exhibit multiple activations, and both cause and prevent one another within a single episode. Therefore, in this paper, we focus on a fully continuous setting involving both multiple relata, generative and preventative relationships under both regular and irregular background conditions.

3 Learning environment

As shown in Figure 2a, the causal devices we investigated in this paper were made up of two "control components" (i.e., Cause A, B) and one "target component" (i.e., Effect E). The connection between each control component and the target component could be generative, preventative, or they could be unrelated. Thus, we focus on learning in a nominal hypothesis space of 9 possible structures (Figure 2f). However, we note that the experimental paradigm and computational models we introduce here can be directly generalized to learning in arbitrarily broader causal hypothesis spaces, as well as under different prior expectations about plausible delays and relations (see below).

We focus on relationships between *point events* (i.e., activations) occurring at a device's components at particular moments in time. As mentioned above, preventative inference is only possible when there is something to prevent [1, 10], e.g., when an effect component has a non-zero chance of occurring in the absence of the preventative influence. Here we thus focus on a setting in which the relevant target component does occasionally spontaneously activate and we contrast two patterns for these base rate activations: (1) moderately *regular* (5 ± 0.5 s, sampled from Gamma$(100, 20)$, Figure 1a) vs. (2) perfectly irregular (5 ± 5 s, sampled from Gamma$(1, 0.2)$, Figure 1b). Meanwhile, an activation of a generative component will *always* produce an "extra" activation of the target component (i.e., the causal power equals 1 [1], Figure 1c), while an activation of a preventative component will block any activations of the target component for a short time window of gamma

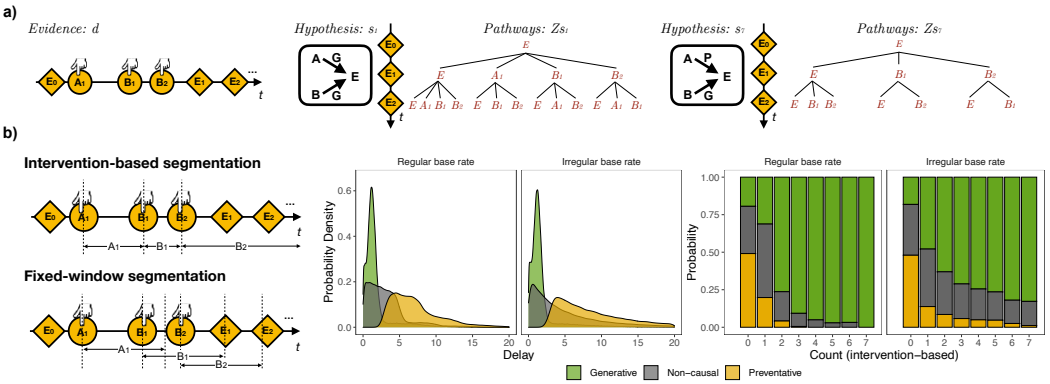

Figure 3: Illustrations for model algorithms. a) Causal path constructions under the normative solution. Circles indicate cause events and diamonds indicate effect events in the evidence. The ideal observer builds different possible pathways (each branch) according to the evidence and hypothetical structures. b) Simulation-and-summary solution. Evidence is segmented under different approaches and learners update beliefs according to statistical distributions of two cues given pre-simulated data.

distributed duration (Figure 1d), irrespective of whether those 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 [21]. Activations of non-causal components have no impact on the behavior of the target component. In our experiments, we assumed generative delays were sampled from Gamma$(9, 6)$ (i.e., 1.5$\pm$0.5 s), and preventative blocking windows were sampled from a Gamma$(36, 12)$ (i.e., 3$\pm$0.5 s). These were chosen simply as an illustrative setting in which 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 actual sampled values for any actual influence are unknown to the learner (human or model), but for simplicity we trained participants on the typical patterns of base rate activations and on typical generative delays and preventative durations in the instructions and so assumed these were available to the model.

In our learning task, participants observed a range of 20-second clips, each showing an abstract causal device's patterns of activation over time resulting from a set of 6 pre-chosen and fixed interventions on control components A and B. Participants then judged which of the 9 potential structure hypotheses best described each device. Each clip began with a base rate activation of the target component then included three interventions on A and three on B randomly spaced and intermingled over 20 seconds. After twenty seconds the clip would end and no further activations could be observed. Figure 3a shows an example timeline for a possible subsequence of around 5 seconds in which there is a single intervention on A, two on B, and three occurrences of E.

4 Modeling continuous-time causal inference

We develop both a normative benchmark for inference and an approximation scheme that may be more psychologically plausible.

We write the effect data (E's activations) as $\mathbf{d}\{d^{(1)}, \ldots, d^{(n)}\}$ indexed in chronological order, conditioned upon a set of intervention $\mathbf{i}\{i_X^{(1)}, \ldots, i_X^{(m)}\}$ with X indicating the activated cause (A or B). The learner then updates the prior over structures $P(S)$ (here assumed as uniform), with a likelihood function $p(\mathbf{d}|S, \mathbf{w}; \mathbf{i})$ to obtain a posterior distribution $P(S|\mathbf{d}, \mathbf{w}; \mathbf{i})$, given the set of gamma parameters $\mathbf{w}$:

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

The next two sections will illustrate two different ways of calculating likelihood $p(\mathbf{d}|S, \mathbf{w}; \mathbf{i})$.

4.1 Normative solution

Normative likelihood calculation depends on an enumerative actual causal attribution step [22]. The basic idea is that the 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 is 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 was activated at 100 ms and B was activated at 1200 ms ($\mathbf{i}\{i_A^{(1)} = 100ms, i_B^{(1)} = 1200ms\}$), and the learner observed two subsequent effects ($\mathbf{d}\{d^{(1)} = 1500ms, d^{(2)} = 2800ms\}$), even assuming 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 links could have failed and meaning either or both 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 $\mathbf{Z}_s$ that could describe what actually happened given each possible structural hypothesis $s \in \mathbf{S}$, summing up the individual likelihood of these possibilities to assess the overall likelihood:

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

Building each possible pathway z follows two steps:

1. Explaining each effect that has been observed.
2. Explaining away any effects that might have occurred but were not observed (for instance that may be still to occur at the end of the observation, or that were prevented from occurring).

Figure 3a shows two examples of the tree of possible causal attributions under two of the possible structural hypotheses. Details of calculations based on gamma distributions are provided in Appendix A.

Since one must consider all possible structure- and all time-consistent cause–effect combinations, the complexity of this inference scheme scales in a worse-than-polynomial manner as the number of events a learner observes increases. This normative inference procedure shows how to solve the current causal learning task in principle. However, in the next section we will introduce a less-demanding method to approximate the normative solution that we will argue provides a more plausible account of how bounded human learners solve the task.

4.2 Simulation-and-summary approximation

Mental simulation has recently been hypothesized to play various important roles in inference [23, 24], while abstraction of complex inputs into useful features and cues is a general principle across learning and decision making. Inspired by this, we consider one way that simulations and abstraction might play into inferences in the current setting. Ullman et al. [7] model inferences about the latent properties of physical objects (such as masses and forces) from observed dynamics using a use a simulation–reality comparison process. They argue that what people compare is not precise spatiotemporal similarity between observations and any single mental simulation but rather the indirect comparison of summary statistics extracted from the detailed observations and simulations under different ground truths. As a simple example, if imagined heavy objects tend to move more slowly than imagined light objects, this could license the use of speed as a (fallible) cue to heaviness. We adopt a similar approach to our causal learning task, exploring whether there are simple local features of event sequences that are diagnostic (if fallible) guides to local causal relationships. The implied cognitive process is that learners draw on (imagined) evidence under different causal ground truth structures develop statistical cues that can be directly applied to pairwise causal judgments within a large causal network. Here we investigate two very simple cues that people might be sensitive to in the current task:

1. Delay: the delay between the control component's activation and the first subsequent target component's activation.

2. Count: the number of the following target component's activations after the control component's activation.

These cues are far from exhaustive their utility is somewhat context-specific, but they are simple to track and turn out to discriminate reasonably well between different causal link types (see Figure 3b). Moreover, they are straightforward to interpret qualitatively. For the delay cue, we would 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 variable interval if there is no connection or a longer interval if it is preventative. For the count cue, more than one effect activation is likely to follow the activation of generative component, which results from the existence of base rate activations, while frequently, no effect activations will follow the activation of preventative components before the next control component event. The former cue considers concrete delay information but ignores the possibility of different causal pathways, while the latter cue also ignores the exact temporal interval between events.

Segmentation approach Given tight limits on human working memory, it is implausible participants would utilize cues based on the full event sequence. Rather, people are found to often use local (i.e., recent) information to make causal inferences [3]. Therefore, we assume that people segment the observation as it unfolds, using recent events to update their beliefs and then discarding their memory of them. A learners' posterior belief distribution is thus conceived as being shaped throughout each clip by a rolling revision. There are at least two ways to segment continuous-time evidence (Figure 3b). A unit of evidence under both approaches begins with the intervention (i.e., the activation of a control component), capturing the basic principle that causes can only influence what happens later. An *Intervention-based* approach thus 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 also ignores actual effects that may not been able to reveal before the occurrence of the next intervention. A *Fixed-window* approach ends one unit of observation after a fixed time window. This has the advantage of stability in its chance of including all relevant effects' but opens the door to confounding influences of multiple interventions occurring within the window. This also depends on some parallel processing since different fixed post-intervention windows are likely to overlap in the timeline. We chose 4-second as a principled fixed-window length in our analysis on the grounds that this is is long enough for different types of links to reveal their influence. We additionally conceive of these cues as modular and independent, i.e., treating the possible influence of the other causal components as noise [25]. For more details, refer to Appendix B.

5 Experiments

We conducted two exploratory experiments with human participants to test how reliably they could infer the structure of causal devices, under different ground truth and delay conditions. Experiment 2 was a replication of Experiment 1 with different instructions and an expanded range of observation sequences.

5.1 Methods

Participants A total of 310 participants (regular vs. irregular: 93 vs. 94 in Experiment 1 and 63 vs. 60 in Experiment 2) were recruited via Amazon Mechanical Turk. They were paid between \$1.00 and \$2.08 depending on their performance.

Design & Procedure The base rate setting was manipulated between subjects with one group experiencing a semi-reliable base rate (5 ± 0.5 s, Gamma$(100, 20)$) and others a fully random one (5 ± 5 s, Gamma$(1, 0.2)$. Participants observed 18 clips in random order. In each case were asked to judge causal structures by selecting among the 9 possibilities (Figure 2d). In Experiment 1, all stimuli were generated by simulating from the ground truth structure. Participants faced two clips for each of the nine structures. As such, the normative model was always more or less confident in the true ground truth structure while the heuristics were more fallible. Therefore, to make a more extensive comparison between models, half the trials in Experiment 2 were generated without a particular ground truth, simply by sampling six interventions and between 1 and 9 effect events in 20 seconds and picked sets of stimuli that differed in their dominant answers under the normative model. We

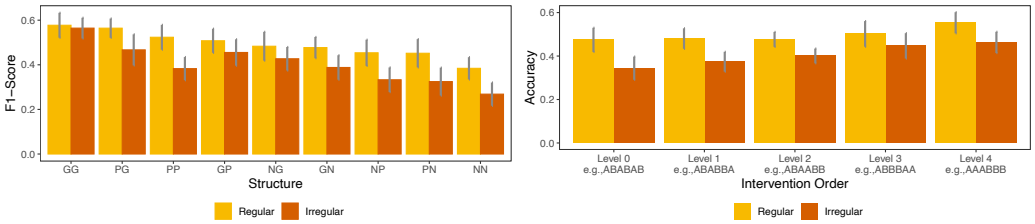

Figure 4: Experiment results with a left panel of F1-score in different structures (ordered by the performance in the regular condition) and right panel of accuracy under different observed intervention orders. Error bars indicate 95% confident intervals.

used a Latin square design to simulate and allocate stimuli, by firstly generate stimuli seeds and then generate stimuli of different causal structures from each seed (See details at Appendix C).

In the experiment interface, component activations were visualized by the requisite component lighting up yellow for 350 ms. The activation of control components was accompanied by a hand symbol and participants were told that this meant the control components were intervened on by someone else, which occurred at random moments (Figure 2b). Participants clicked a "Start" button to watch the clip in each trial, and then marked their answers for two connections during or after the clip by clicking the connection (Figure 2d). Each clip could only be played once.

In Experiment 1, participants were told about and trained on the timing of three types of connections as well as the target component's self-activation prior to the inference task. To help participants understand the regular vs. irregular base rate conditions, we used two different metaphors (Figure 2e). Participants on the regular condition saw an illustration with a circular sign to enforce the impression of a roughly periodic activation, whereas participants on the irregular condition saw an illustration with an exogenous link to the cue that something sometimes activates the target component directly but one cannot anticipate when this will next happen. Participants also completed one practice trial with a causal device that included one generative connection and one non-causal connection. Feedback was provided in the practice trials but not the test trials. However, the video training on timings and the practice trial actually provide a form of "labeled data" to participants. Since this slightly confound with the assumption of the simulation-and-summary approach that statistical cues emerge from learners own mental simulation, we opted to remove the video training and practice trials in Experiment 2. To properly incentivize judgments, we paid a 3-cent bonus for each correctly identified connection during the main experiment in addition to the basic $1 payment.

5.2 Results

We first analyze participants' accuracy by ground truth (focusing on all clips in Experiment 1 and the stimuli in Experiment 2 that had a ground truth). We investigate whether participants' performance was influenced by base rate regularity, causal structures, and observed intervention orders.

Accuracy The accuracy for each participant per *connection* was significantly above chance (33%) in both the regular condition (EXP1: 66%$\pm$ 22%, $t(92) = 14.89, p < .001$; EXP2: 67%$\pm$22%, $t(62) = 12.05, p < .001$) and irregular condition (EXP1: 61%$\pm$ 18%, $t(93) = 14.84, p < .001$; EXP2: 59%$\pm$19%, $t(59) = 10.24, p < .001$). Accuracy was above chance for all three connection types taken separately in both regular and irregular conditions ($ps < .001$). The accuracy for each participant at the *structure* level (1 = correct in both connections; 0 = otherwise) was again substantially higher than chance (11%) in the regular condition (EXP1: 49%$\pm$27%, $t(92) = 13.87, p < .001$; EXP2: 49%$\pm$ 29%, $t(62) = 10.67, p < .001$) and the irregular condition (EXP1: 41%$\pm$22%, $t(93) = 13.32, p < .001$; EXP2: 39%$\pm$23%, $t(59) = 9.24, p < .001$).

We used logistic mixed-effect analyses to compare participants' performance under different levels of base rate regularity. At the connection level, with subjects, stimulus seeds, and connection types (generative, non-causal, preventative) as random factors, the accuracy in the regular group was higher than the accuracy in the irregular group (EXP1: $\beta = 0.31, z = 2.07, p = .04$; EXP2: $\beta = 0.45, z = 2.22, p = .03$). At the structure level, with subjects, stimulus seeds from Latin square design, and ground truth structures as random factors, the accuracy in the regular group was

Table 1: Model accuracy and fitting results

	ACC	Parameters	BIC	CV	BestN (BIC)	BestN (CV)
Experiment 1						
Normative	93%/70%	$\tau : 0.44$	12170	-6084	31	34
Simulation-and-Summary (intervention-based):					(84)	(94)
Delay	60%/58%	$\tau : 0.31$	11990	-5994	26	24
Count	42%/32%	$\tau : 0.2$	12134	-6065	47	45
Combined		$\tau_d : 0.50$; $\tau_c : 0.39$	**11719**	**-5855**	11	25
Simulation-and-Summary (fixed-window):					(35)	(50)
Delay	65%/63%	$\tau : 0.35$	12084	-6040	16	15
Count	58%/46%	$\tau : 0.33$	12461	-6228	17	19
Combined		$\tau_d : 0.48$;$\tau_c : 0.87$	11987	-5990	2	16
Random	11%/11%		14859	-7430	37	9
Experiment 2						
Normative	95%/69%	$\tau : 0.57$	8845	-4433	13	13
Simulation-and-Summary (intervention-based):					(65)	(70)
Delay	68%/62%	$\tau : 0.33$	8277	-4136	20	20
Count	40%/36%	$\tau : 0.20$	8343	-4173	38	34
Combined		$\tau_d : 0.55$; $\tau_c : 0.36$	**8111**	**-4053**	7	16
Simulation-and-Summary (fixed-window):					(27)	(37)
Delay	74%/64%	$\tau : 0.39$	8393	-4196	15	17
Count	63%/43%	$\tau : 0.36$	8585	-4292	9	9
Combined		$\tau_d : 0.54$; $\tau_c : 0.90$	8336	-4166	3	11
Random	11%/11%		9774	-4887	18	3

Note: Model accuracy (in identifying the ground truth structure) was calculated prior to the fitting of human data under regular/irregular conditions separately.

also higher than the accuracy in the irregular group (EXP1: $\beta = 0.44, z = 2.35, p = .02$; EXP2: $\beta = 0.55, z = 2.19, p = .03$).

Structure differences We used F1-score as an index to compare participant's performance in detecting different causal structures. Because F1-scores for a single participant would be invalid if they never chose a certain structure, we calculated F1-scores based on groups of clips that used the same seed. This led to 27 data points (18 in Experiment 1 and 9 in Experiment 2) for each type of structure in regular as well as irregular conditions. As shown in Figure 4, participants were best at identifying structures that contain two generative links (GG) and worst at detecting structures that were made of two non-causal links (NN). They generally performed better in structures without non-causal links (GG, GP, PG, PP).

Observed intervention orders Finally, we explored whether intervention order affected performance. This was considered due to the theoretical deviation between holistic vs. modular thinking in normative vs. simulation-and-summary models. We classify different intervention orders according to the extent that one control component is intervened repeatedly. The most repetitive pattern is AAABBB or BBBAAA (Level 4, the mirrored orders that begin with B are omitted in later demonstration), and the least repetitive is ABABAB (Level 0). As shown in Figure 4, participants' accuracy increases as the intervention orders become more repetitive, for both regular ($\beta = 0.11, z = 2.05, p = .04$) and irregular ($\beta = 0.16, z = 3.22, p = .001$) base rate, after controlling for the subject, ground truth structure, and even the exact intervention timings (i.e., the seeds).

6 Model fitting

We fit our normative and heuristic models to both aggregate and individual structure judgments (including stimuli with or without the ground truth answers). We assumed that participants selected their responses according to a softmax rule controlled by "temperature" parametersτ [26] (See Appendix D).

We evaluate model fit using both cross-validated log likelihood based on leaving out trials from each seed, and also using BIC. For both cross-validation and BIC, participants were better fit by the simulation-and-summary approach that combines both the "delay" and "count" cues, and segments evidence according to intervention actions (Table 1). At the individual levels, more participants were best fit by one of the simulation-and-summary models than by the normative model or the random baseline. We also find some patterns that reflect the results of F1-scores and intervention order analyses (regarding stimuli with ground truths). For example, although participants were worst at identifying structures with non-causal links, the normative learner was less certain for structures with two preventative links (PP) (Figure E.1). Also, normative performance was not very affected by the intervention order while human and the heuristics' performance was a bit more sensitive to this, especially for the regular base rate condition (Figure E.2).

We additionally performed a grid search in $[2, 2.5, 3, 3.5, 4, 4.5, 5]$ seconds to find whether any specific fixed-window length beats the inter-intervention-window-based approach. As shown in Figure E.3, models with different fixed-window lengths always had substantially larger BICs than the model with the inter-intervention window approach. This was true despite the fact that the models' accuracy in causal identification is quite sensitive to the window length.

7 Discussion

In this paper, we explored algorithms for learning causal structure from point events occurring in continuous time, and compared these to how human learners perform in this setting. While classical causal cognition research has focused on contingency data and generative relationships, this is some of the first work that shows that people can learn successfully in situations where causal systems involving both generative and preventative relationships interact in shaping event sequences over time.

In terms of modeling, we for the first time introduced considerations of prevention into an "actual causal attribution" process model [22]. By exhaustively constructing possible actual causal paths to explain observed data, our normative model demonstrates that near-perfect performance is possible in this setting, at least given the correct delay assumptions and unlimited processing power. Our normative models had higher accuracy than both participants and our simulation-and-summary models, indicating that actual attribution, the top rung of Pearl's so-called "ladder of causation" [27], is key for achieving benchmark levels of accuracy. The inference and approximation we present in this paper while context-specific in its particulars, is very broad in its principles, and is not restricted to the current setting but can be modified to handle a wide range of inferences about causal systems. Essentially, any system can be represented with a causal mechanism that produces point events over time that can be inferred and reasoned about in this way.

Despite the accuracy win for the normative approach, human responses were better captured by the simulation-and-summary model. This involves identifying locally diagnostic summary statistics that can be used to make approximate local-structure inferences. This approximation could be further specified by having different assumptions about what statistical cues people use to make comparisons efficiently, and how people segment the evidence during an ongoing observation. We explored two heuristic cues: delay and count, which assumed that people track the delays between putative cause-effect activations or else count the number of effect events between each putative cause event. They then compare observed patterns to the patterns characteristic to each edge hypothesis. The approach sacrifices precision in terms of actual causal attribution but may capture how people manage the information stream in real-time casual induction, given the basic cognitive limits on real-time processing [28]. This also demonstrates one possible way that mental simulation could contribute to temporal causal reasoning, as an extension to current perspectives on the role of mental simulation in physical reasoning [29, 7]. The adequacy of this local summary information for pairwise causal attribution is likely to depend on the particulars of the setting (i.e., how sparse the relationships are,

how reliable the delays, how much background noise and so on). However, provided there is at least moderate stability in these dimensions, there is space for such local heuristic cues like the ones we advance here to emerge and be exploited by bounded human learners.

As for how people segment long observation in continuous time, we investigated two approaches: based on the intervention actions or based on time windows with fixed length. It turned out that the intervention-based approaches fit participants' performance better here, although the fixed-window approaches had higher accuracy. We take this to suggest that people may struggle to monitor causal delays from multiple perspectives in parallel. That is, they appear to absorb evidence, update their local hypotheses and forget the evidence that has been "used" before the next intervention.

In sum, this paper investigated causal structure induction from observation of real-time event patterns involving prevention as well as generation. Participants were capable of real-time causal structure induction in this setting and our modeling suggested they may achieve this via statistical cues such as average delays and counts that are much easier to track in real time than the exact generative model likelihoods. We also assumed that learners parse evidence in an online learning setting by segmenting it into chunks, in this case between each intervention, using only the latest chunk to update their beliefs and doing so modularly, focusing on one pair of components at a time. This work thus provides a quantitative sketch of how human learners succeed in identifying causal structure from temporal dynamics. It could contribute to our understanding of natural cognition and sheds light on the question of how any cognitive agent can make causal inferences in continuous natural environments.

References

[1] Patricia W Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367, 1997.

[2] Thomas L Griffiths and Joshua B Tenenbaum. Structure and strength in causal induction. *Cognitive Psychology*, 51(4):334–384, 2005.

[3] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3):301, 2017.

[4] Neil R Bramley, Tobias Gerstenberg, Ralf Mayrhofer, and David A Lagnado. Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(12):1880–1910, 2018.

[5] Neil R Bramley, Tobias Gerstenberg, Ralf Mayrhofer, and David A Lagnado. Intervening in time. *Time and causality across the sciences*, pages 86–115, 2019.

[6] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, 1982.

[7] Tomer D Ullman, Andreas Stuhlmüller, Noah D Goodman, and Joshua B Tenenbaum. Learning physical parameters from dynamic scenes. *Cognitive Psychology*, 104:57–82, 2018.

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

[9] Marc J Buehner, Patricia W Cheng, and Deborah Clifford. From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1119, 2003.

[10] Melissa Wu and Patricia W Cheng. 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, 1999.

[11] Judea Pearl. *Causality*. Cambridge University Press (2009 reprint), New York, 2000.

[12] Anna Coenen, Bob Rehder, and Todd M Gureckis. Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, 79:102–133, 2015.

[13] Benjamin M Rottman and Frank C Keil. Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64(1-2):93–125, 2012.

[14] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

[15] David R Shanks, Susan M Pearson, and Anthony Dickinson. Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, 41(2):139–159, 1989.

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

[17] W James Greville and Marc J Buehner. Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, 139(4):756–771, 2010.

[18] Michael Pacer and Thomas L Griffiths. Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 833–838, 2012.

[19] Simon Valentin, Neil R Bramley, and Christopher G Lucas. Learning hidden causal structure from temporal data. In *Proceedings of the 42th Annual Conference of the Cognitive Science Society*, pages 1906–1912, 2020.

[20] Simon Stephan, Ralf Mayrhofer, and Michael R Waldmann. Time and singular causation–a computational model. *Cognitive Science*, 44(7):e12871, 2020.

[21] Christopher Carroll and Patricia Cheng. Preventative scope in causation. In N. A. Taatgen and H. van Rijn, editors, *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, pages 833–838, 2009.

[22] J Y Halpern. *Actual causation*. MIT Press, 2016.

[23] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[24] Tobias Gerstenberg and Joshua B Tenenbaum. Intuitive theories. In M. Waldmann, editor, *The Oxford handbook of causal reasoning*, pages 515–548. Oxford University Press, New York, 2017.

[25] Philip M Fernbach and Steven A Sloman. Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):678, 2009.

[26] R Duncan Luce. *Individual choice behavior*. Wiley, Hoboken, 1959.

[27] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, New York, 2018.

[28] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:1–60, 2020.

[29] Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 2021.

A Normative calculations

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. The Step 1 and 2 correspond to path construction in the main text. 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.

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:

$$P(g' \rightarrow e'|\alpha_g, \beta_g) = P(t_{g' \rightarrow e'} = t_{g'e'}|\alpha_g, \beta_g) \tag{A.1}$$

Step 2 involves forming $g' \rightarrow h$ pairs where h is a hidden effect event assumed to happen some time 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:

$$\begin{aligned} P(g' \rightarrow h|\alpha_g, \beta_g, \alpha_p, \beta_p) = P(t_{g' \rightarrow h} > t_{end}|\alpha_g, \beta_g) + \\ P(t_{g' \rightarrow h} t_{end}|\alpha_g, \beta_g) \prod_{p'} P(t_{g' \rightarrow h} < t_{g'h} + t_{p' \rightarrow h}|\alpha_g, \beta_g, \alpha_p, \beta_p) \end{aligned} \tag{A.2}$$

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 $g' \rightarrow h^{(1)} \rightarrow \ldots \rightarrow h^{(n)} \rightarrow e'$, where $h^{(1)} \ldots h^{(n)}$ happens within the preventative windows. Meanwhile, according to the summing property the gamma distribution, if $X, Y \sim Gamma(\alpha, \beta)$ then $X + Y \sim Gamma(2\alpha, \beta)$. The probability $P(g' \rightarrow h^{(1)} \rightarrow \ldots \rightarrow h^{(n)} \rightarrow e')$ can thus be represented as Eq. A.3, where the calculation of $P(g' \rightarrow e')$ is similar to Eq. A.1, and the calculation of $P(g' \rightarrow h^{(n')})$ is similar to Eq. A.2 except that t_{end} is substituted with $t_{e'}$ and only the second item of prevention is considered.

$$\begin{aligned} P(g' \rightarrow h^{(1)} \rightarrow \ldots \rightarrow h^{(n)} \rightarrow e'|\alpha_b, \beta_b, \alpha_p, \beta_p) = \\ P(g' \rightarrow e'|(n+1) \cdot \alpha_b, \beta_b) \prod_{n' \in n} P(g' \rightarrow h^{(n')}|n \cdot \alpha_b, \beta_b, \alpha_p, \beta_p) \end{aligned} \tag{A.3}$$

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 gamma cumulative density function of prevention:

$$P(p' \rightarrow e'|\alpha_p, \beta_p) = P(t_{p' \rightarrow e'} < t_{p'e'}|\alpha_p, \beta_p) \tag{A.4}$$

B Simulation-and-summary calculations

Characteristic summary statistics for each structure hypothesis were constructed by simulating 10,000 sequences of point events from each structure type, with three interventions on A or B, and then calculating the empirical features for each intervention in each structure. This results in 60,000 simulated cases. Distinct from the experimental stimuli, simulated sequences here were not cut at twenty seconds so as to avoid the complex boundary effect in distribution constructions. By its definition we can see that the delay cue is independent of segmentation approaches since it always relates to the nearest effect event, while the count cue is sensitive on the segmentation for which we need to build distributions for intervention-based and fixed-window assumptions separately. Delay distributions use the probability density function smoothed with Gaussian kernels, and Count distributions used the discrete probability mass functions directly. When observing a new interventions, the probability of each causal structure was estimated by the normalized posterior of the summary statistic calculated on the observed data.

Inherent to this heuristic approach is the radical simplifying assumption that the features of the evidence subsequent to each control component event are modular and independent, that is, that

one can safely ignore that the subsequent device behavior also depends on the behavior of the other control component(s). Thus, each connection was estimated independently as generative, non-causal, or preventative, and then combined to yield a probability for each causal structure. For example, an intervention on A with a nearest effect occurring 2.5 seconds later has a posterior of [.2, .7, .1] of having being 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 (under the assumption of uniform prior distributions). When the next intervention on A happens, the likelihood will be updated by combining the new probability with the original one.

The boundary situations we considered were as follows: If no effect happens within the observation window, in both segmentation approaches, the delay cue will be marked as larger than the observing window and the probability will be estimated according to cumulative density function. If the observation window is less than the designed window length in the fixed-window approach (which often happens near the end of the clip), or there is no next intervention in the intervention-based approach, the count cue will be marked as greater than or equal to the observed count of effects and the probability will also be estimated on the basis of cumulative mass functions.

C Experiment stimuli generation and allocation

To ensure participants' performance on different conditions were comparable, the stimuli generation and assignment procedure was as follows: In Experiment 1, eighteen seeds were created independently. Each of them included a set of timings of interventions, regular base rate activations, irregular base rate activations, and what generative delays (or blocking windows) A and B would have if they were generative (or preventative) components. Then under each seed, 18 stimuli (9 causal structures $\times$ 2 base rate settings) were generated by implementing generative or preventative influences according to the grounded structure. All stimuli were finally divided into 18 sets (9 sets for each base rate setting) according to the Latin-square design that ensured participants would only see only one structure under each seed. Participants were randomly assigned to one of 18 sets. The half of the stimuli in Experiment 2 that have ground-truth answers also followed the procedure above.

D Softmax rules

We assumed that participants selected their response according to a softmax over a posterior value vector v:

$$P(n) = \frac{\exp(v_n/\tau)}{\sum_{n' \in N} \exp(v_{n'}/\tau)} \tag{D.1}$$

The "temperature" parameter $\tau \in (0, +\infty]$ controls how consistent the participant is in selecting the answer with the largest v_n in choice n. Smaller τ means that the participant's answer is better aligned with the model's answer with τ approaching $+\infty$ modeling random selection. For the normative model we simply set v_n to $P(s|\mathbf{d}, \mathbf{w})_n$, as well as the single cue models in the stimulation-and-summary approach. For the combination of two cues, we use two temperatures τ_d and τ_c to give weights to the delay and count cues:

$$P(n) = \frac{\exp(v_{dn}/\tau_d + v_{cn}/\tau_c)}{\sum_{n' \in N} \exp(v_{dn'}/\tau_d + v_{cn'}/\tau_c)} \tag{D.2}$$

E Model Performance

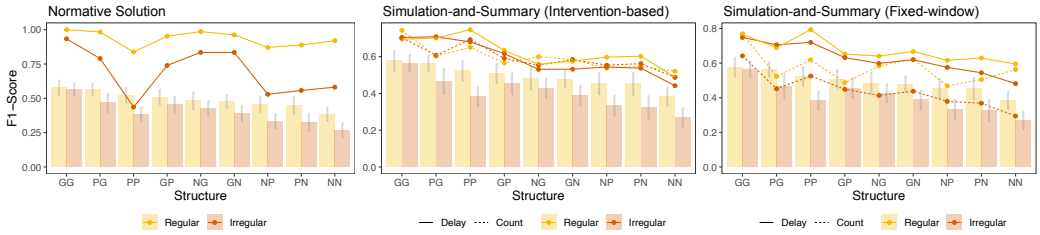

Figure E.1: Models' F1-score under different structures of experimental stimuli. Bars in the background indicate human performance.

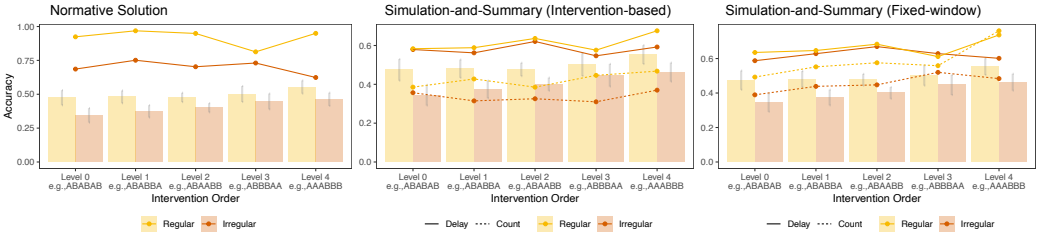

Figure E.2: Models' judgment accuracy under different intervention orders of experimental stimuli. Bars in the background indicate human performance.

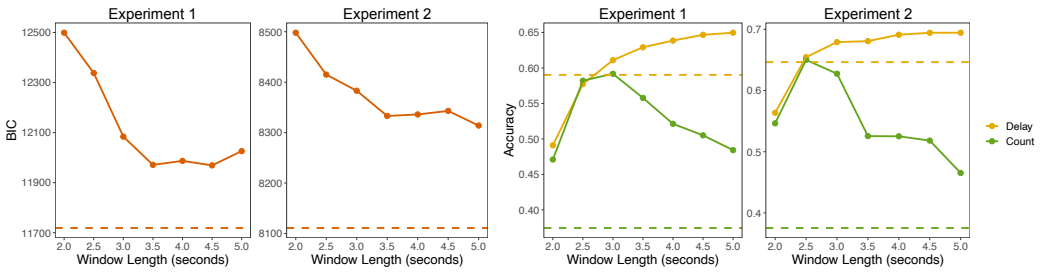

Figure E.3: BIC and model accuracy under different fixed window lengths of simulation-and-summary models. Horizontal dashed lines indicate cases of intervention-based segmentation.