

Causal Structure Learning in Continuous Systems

Zachary J. Davis¹, Neil R. Bramley², and Bob Rehder¹

¹Department of Psychology, New York University

²Department of Psychology, The University of Edinburgh
Affiliation

Abstract

Causal systems are complicated. Despite this, causal learning research has traditionally emphasized how causal relations can be induced on the basis of idealized events, i.e. those that have been mapped to binary variables and abstracted from time. For example, participants may be asked to assess the efficacy of a headache-relief pill on the basis of multiple patients who take the pill (or not) and find their headache relieved (or not). In contrast, the current study examines learning via interactions with continuous dynamic systems, systems that include continuous variables that interact over time (and that can be continuously observed in real time by the learner). To explore such systems, we develop a new framework that represents a causal system as a network of stationary Gauss–Markov (“Ornstein–Uhlenbeck”) processes and show how such *OU networks* can express complex dynamic phenomena such as feedback loops and oscillations. To assess adult’s abilities to learn such systems, we conducted an experiment in which participants were asked to identify the causal relationships of a number of OU networks, potentially carrying out multiple, temporally-extended interventions. We compared their judgments to a normative model for learning OU networks as well as a range of alternative and heuristic learning models from the literature. We found that, although participants exhibited substantial learning of such systems, they committed certain systematic errors. These successes and failures were best accounted for by a model that describes people as focusing on pairs of variables, rather than evaluating the evidence with respect to the full space of possible structural models. We argue that our approach provides both a principled framework for exploring the space of dynamic learning environments as well as new algorithmic insights into how people interact successfully with a continuous causal world.

Introduction

We live and act in a messy world. Scientists’ best models of real-world causal processes typically involve not just stochasticity, but real-valued variables, complex functional forms,

delays, dose-dependence, and feedback leading to rich and often nonlinear emergent dynamics (Cartwright, 2004; Sloman & Lagnado, 2015; Strevens, 2013). It follows that learning successfully in natural settings depends on accommodating these factors. Cognitive psychologists have explored many of these dimensions of complexity in isolation (e.g. *stochasticity*: Bramley, Dayan, Griffiths, & Lagnado, 2017; Rothe, Devereitt, Mayrhofer, & Kemp, 2018; Waldmann & Holyoak, 1992; *interventions*: Coenen, Rehder, & Gureckis, 2015; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005; *?*; *time*: Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Buehner & May, 2003; Lagnado & Sloman, 2006; Rottman & Keil, 2012; and *continuous variables*: Pacer & Griffiths, 2011). However, we argue these components generally can not be isolated in realistic learning settings, meaning a deeper understanding of human causal cognition will require a new framework that naturally accommodates inference from interventions in continuous dynamic settings.

As an everyday example of a time-sensitive, dose-dependent causal relationship, consider the complexities involved in consuming alcohol. It is common for drinkers to adjust their consumption based on their recognition that higher doses affect inhibition or mental clarity, that will in turn have other downstream effects on quality of conversation or willingness to sing karaoke. The effects of alcohol consumption differ widely in quality and quantity depending on dosage and time delays. A small glass of wine with dinner will likely have little effect on mental clarity whereas a few shots will have a stronger effect. Further complicating the learning problem, these effects of alcohol do not come instantaneously but are rather delayed and distributed in time. Worse still, more complex dynamics exist, such as the feedback loop between lowered inhibition and increased alcohol consumption. And there are innumerable contributing factors such as diet or amount of sleep that can modulate alcohol's effect. Thus in settings like this, the learning problem is non-discrete (how much alcohol did I drink) and extended in time (when did I drink it), produces evidence that is naturally time ordered (how you feel over the preceding and subsequent hours), and contains complicated dynamics (e.g. feedback loops). In the current paper, we study human learning through real-time interactions with causal systems made up of continuous valued variables. We see this setting as capturing the richness of real world causal learning, while remaining simple and principled enough to allow for close formal analysis.

The structure of the paper is as follows. First, we summarize relevant past work on causal structure inference from interventions, temporal information, and different representations of functional form. Next, we lay out our new formalism for inference of causal structure between continuous variables. We then report on an experiment, in which participants interact with causal systems represented by sliders on the computer screen. We provide an exploratory analysis of the interventional strategies we observed in the experiment before analyzing structure learning through the lens of a normative Bayesian inference model and a range of heuristic and approximate alternatives, finding evidence that people focus sequentially on individual connections rather than attempting to learn across the full space of possible causal models at once. Finally we discuss new opportunities provided by the formalism introduced in this paper, including future questions in causal cognition as well as applications to other areas such as dynamic control.

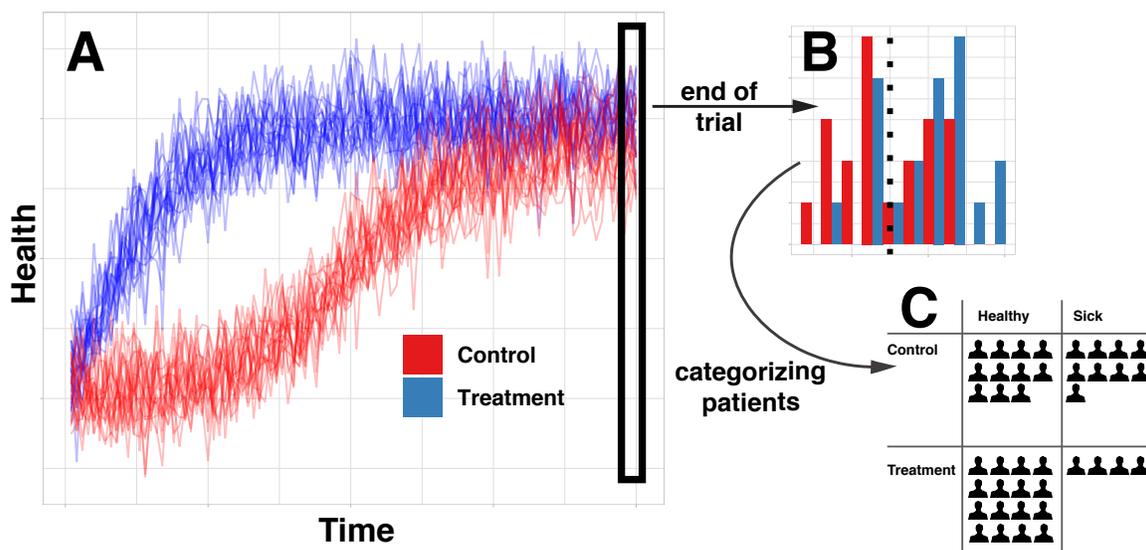


Figure 1. Illustration of abstraction from full timeseries data to probabilistic contingency. (A) is a full time course of the health of 40 simulated patients throughout the course of a classic randomized controlled trial. (B) demonstrates the type of information available when only evaluating the health of patients at the end of the trial. (C) demonstrates the type of information available when categorizing patients into ‘sick’ and ‘healthy’ groups, rather than maintaining full continuous information.

Past research

Probabilistic causation over discrete events. Research in causal cognition has generally aligned itself with the philosophical tradition of probabilistic causation, which defines a causal relationship as one where a cause changes the probability of its effect (Hitchcock, 1997). This definition implicitly operates over particular representations: discrete states such as events or facts that have some probability of occurring or being true. Because of this, experimental work in causal cognition has primarily focused on causal relationships between discrete valued (often binary) variables (e.g. Ali, Chater, & Oaksford, 2011; Fernbach & Erb, 2013; Hayes, Hawkins, Newell, Pasqualino, & Rehder, 2014; Krynski & Tenenbaum, 2007; Rehder, 2014; Rothe et al., 2018; Sloman, 2005). These are typically presented in contexts in which temporal information is either unavailable or abstracted away so that cases can be summarized in a contingency table. See Figure 1 for a simple example in which (A) continuous data is (B) snapshotted in time, in order to (C) dichotomize and create counts of contingencies and abstract this into a probabilistic causal relationship. This approach is very common in part because there is a well established mathematical framework—*Bayesian networks*—for efficiently encoding joint distributions of sets of variables in the form of networks of probabilistic contingencies (Barber, 2012; Pearl, 2009).

While the probabilistic contingencies paradigm has been fruitful for exploring many aspects of causal cognition, we are interested in other settings. As mentioned, we believe that many real life systems may not lend themselves to discretisation, nor involve much independent and identically distributed data with no temporal information. Instead, people

are often have access to autocorrelated, time-dependent, continuous information and we are interested in they how represent and draw inferences on the basis of this information.

Learning. A prominent question in causal cognition is how people learn causal relationships from contingency data such as that presented in Figure 1C. Although the literature shows that humans are often quite adept causal learners (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) there are a number of important exceptions. One is that updates to beliefs about causal structure on the basis of new information are often made narrowly rather than optimally, that is, in ways that do not compare the evidential fit against the entire hypothesis space. To model this, Fernbach and Sloman (2009) introduced a *Local Computations* (LC) model, which posits that people focus “on evidence for individual causal relations rather than evidence for fully specified causal structures”. By ignoring the possible influences of other causes in the model, their model captures a strong empirical pattern in which human learners exhibit order effects and tend to overconnect their causal hypotheses (also see Taylor & Ahn, 2012). Bramley, Dayan, et al. (2017) extended this finding to their more unconstrained task, finding evidence that people consider local changes to adapt their previously favoured hypothesis. Together, these studies suggest that people use a local updating strategy, testing and evaluating individual causal links rather than updating the full posterior space. We ask whether this tendency toward local learning extends to the continuous dynamic systems that are under study here.

Learning via interventions. As well as capturing probabilistic relationships, Bayesian networks can be used to reason about, and from, idealized manipulations of causal systems, or “interventions” (Pearl, 2009). Bayesian networks, at their core, deal with *independence*, not dependence, relations. Because of this, if a cognizer passively observes some variables but cannot observe the temporal direction of their influences (i.e. perhaps they influence one another too quickly to see) they can be equally consistent with multiple causal hypotheses. For example, the common cause $X \leftarrow Y \rightarrow Z$ and chain $X \rightarrow Y \rightarrow Z$ are “Markov equivalent” because, in both networks, X and Z are independent conditional on Y . However, crucially, Markov equivalent networks do not have identical data distributions under intervention. In the example of Markov equivalent networks given above, intervening to set Y to some value y as denoted with Pearl’s 2009 “Do()” operator, would change the distribution for X under the common cause — i.e. $P(X) \neq P(X|\text{Do}[Y = y])$ for at least some y — but would not affect the distribution for X for the chain — i.e. $P(X) = P(X|\text{Do}[Y = y])$ for any y .

It has been shown that people are able to learn successfully from interventions, and are often moderately efficient in their intervention selection according to information-optimal norms (Bramley, Dayan, et al., 2017; Coenen et al., 2015; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005). However, participants in these studies also typically exhibited biases indicative of the influence of cognitive constraints. For example, Coenen et al. (2015) found that, when comparing between two potential causal networks, people appeared to follow a heuristic of intervening on nodes to produce as many effects as possible (averaged across the candidate networks) rather than intervening to maximally distinguish between the two. Use of this heuristic was more common when intervening under time pressure. Bramley, Dayan, et al. (2017) tested people’s learning in a broader hypothesis space encompassing all possible 3 and 4 variable network structures. They found that people made interventions that appeared to target uncertainty

about a specific individual link, node or hypothesis, rather than interventions effective at reducing their uncertainty “globally” over all possible causal graphs. Here we assess the efficacy of learners’ interventions on continuous dynamic systems for which variables are potentially manipulated through a range of magnitudes over an extended period of time.

Time. Time has long been seen as a powerful cue for causation (Hume, 1740), especially with regards to identifying causal direction. People rule out backwards causation, assuming that effects cannot precede causes (Bramley, Gerstenberg, & Lagnado, 2014; Burns & McCormack, 2009; Greville & Buehner, 2010). Work in the cognitive sciences on the use of time in causal judgments has focused on point events separated by delays — that is, events like explosions and collisions that occur at particular times but with negligible duration (Griffiths, 2004; Lagnado & Sloman, 2006; McCormack, Frosch, Patrick, & Lagnado, 2015; Pacer & Griffiths, 2012; Shanks, Pearson, & Dickinson, 1989). From this line of work, we have learned more than just that temporal order is relevant for causal direction. The actual temporal dynamics of causal systems affect judgments, for example shorter and more reliable delays between cause and effect are more readily seen as causal (Bramley et al., 2018; Greville & Buehner, 2010).

In the most systematic work on people’s use of temporal dynamics to learn causal structure, Bramley and colleagues combined interventions and time to investigate people’s learning of causal structure between components that exhibited occasional (punctate) events that could also be brought about by interventions (Bramley et al., 2018; Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017). They found that people are sensitive to expected delays, especially when they also expect the true delays to be reliable, and are judicious and systematic in their use of interventions to perturb the system. While these studies have been valuable in demonstrating that people are sensitive to the temporal characteristics of causal systems, many everyday systems — such as economies, ecosystems, or social groups — are more naturally described as extended shifting influences than point events. We extend the study of time on causal cognition to explore these new, unstudied phenomena.

Continuous Variables. As discussed above, many natural scenarios involve continuous valued variables and causal influences that are typically extended in time rather than punctate. Given the ubiquity of such systems, continuous variables have received surprisingly little attention in the study of causal cognition. In a reanalysis of data from Marsh and Ahn (2009) as well as a study of their own, Pacer and Griffiths (2011) showed that people are capable of learning individual cause-effect relationships between continuous variables. Soo and Rottman (2018) investigated causal relations in nonstationary time series, i.e. those where the average value of the variables changes in time in a way unrelated to the causal relations between those variables. They proposed three ways that the variables could be represented before being correlated: (1) state values, (2) difference scores, and (3) trinarized difference scores (positive, negative, or zero). In their task, people’s causal strength judgments were based on the correlations between the discretized *changes* in variables’ values over time rather than correlation between the variables themselves.

In sum, our approach here is novel in two key respects. First, we study a setting that, like reality, is continuous in terms of both time and state space. This allows us to study learning in the context of causal systems that give rise to nonlinear emergent dynamics and resultant evidence patterns that have not previously been explored in the context of human causal inference. Second, we explore an interactive setting in which participants

intervene on the system of interest in complex, extended ways, rather than merely passively observing its behaviour or setting states across discrete trials, again mapping better onto real world interventions.

The task

We chose a simple and intuitive structure learning task that allows for learners to use their mouse to interact with the variables in a system represented by a set of moving sliders on the computer screen. A depiction of how the sliders were presented is shown in Figure 4. Participants could observe the evolving sequence of variable values but also move and hold the variables (one at a time) at positions of their choice by using the mouse. As mentioned, this environment allows us to test learning of causal systems with continuous valued variables and feedback dynamics. It also allows us to assess learning via interventions that are both extended over time (learners choose how long to intervene) and nonstationary (learners might “hold” the variable in a particular position or “wobble” it up and down).

Continuous Causality in Time

This section presents a formalism for modelling causal systems that relate continuous variables in time. To define a generative model for such systems, we first introduce the notion of an Ornstein–Uhlenbeck (OU) process and then define how multiple OU processes can be interrelated so as to form an interacting causal system. We then describe normative inference within this model class on the basis of both observational and interventional data.

Generative model

The Ornstein–Uhlenbeck process. An Ornstein–Uhlenbeck (OU) process is a stationary Gauss-Markov process that reverts to a stable mean (Uhlenbeck & Ornstein, 1930). It can be conceptualized as Brownian motion with the addition of a corrective force that biases the process’s expected value towards the mean of the distribution. The magnitude of that force increases as a function of the distance between that mean and the process’s current state. Formally, Δv_i^t — the change in variable i from time t to $t + 1$ — is defined as

$$P(\Delta v_i^t | \omega, \mu_i, v_i^t, \sigma) = \omega[\mu_i - v_i^t] + N(0, \sigma) \quad (1)$$

where v_i^t is the value of i at time t , μ_i is the mean of the process for variable i , σ is its variance, and ω is a parameter greater than 0 that determines how sharply the process reverts to the mean.¹ μ_i is also referred to as the process’s *attractor state* because it is the value to which the process will revert to at asymptote. See Figure 3A for an example of an OU process with an attractor state of 0.

OU Processes and Causality. This definition can be generalized to accommodate OU processes with nonstationary means. In particular, we take the step of assuming that the attractor state μ for a variable is determined by some function of the most recent values of its cause(s). When a variable has no causes we model its attractor state as being 0.

¹Throughout this work we use subscripts to denote variables and superscripts to denote time. Note that whereas v_i^t is the value of i at time t , v_i is the value of i at all timesteps, v^t is the value of all variables at time t , and v is the value of all variables at all times.

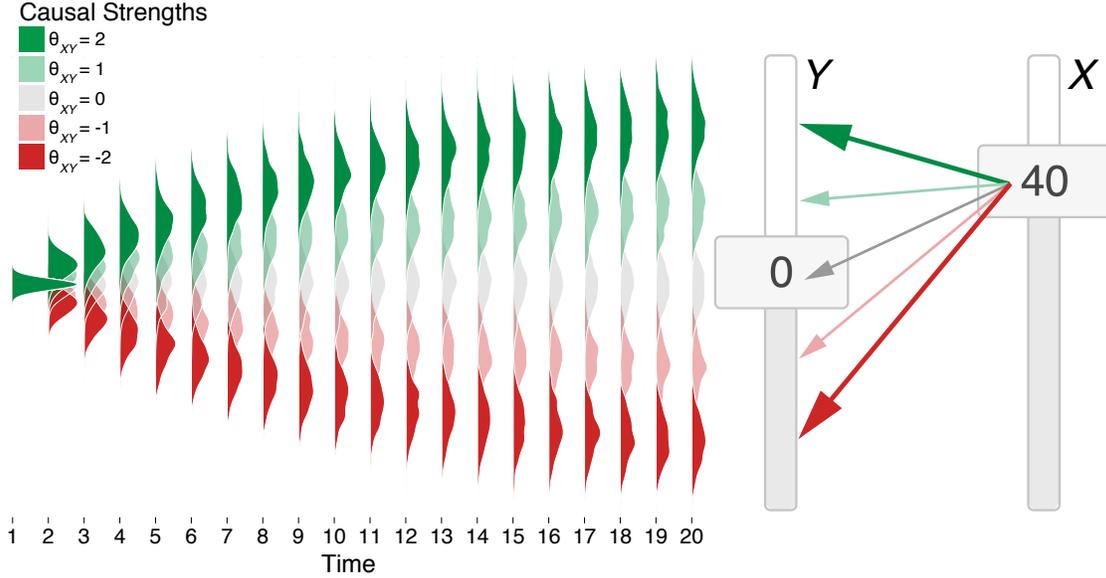


Figure 2. Visualization of the impact of a single cause (slider X) on a single effect (slider Y) in an OU network with different causal strengths. Slider X is held to a value of 40 for 20 timepoints, leading slider Y to unfold over time to different values depending on the causal strength. Probability distributions are smoothed averages of 100 runs of the network given different causal “strengths” θ_{XY} (colored shading) where $\omega = .1$ and $\sigma = 5$.

The single cause case. For a variable i with a single cause j this function is simply,

$$\mu_i^{t+1} = f(v_j^t) \quad (2)$$

where v_j^t is the value of j at time t . As j changes over time, so too does the output of $f(v_j^t)$, which serves as the new attractor state of variable i at the next timepoint. For simplicity, here we assume that $f(v_j^t)$ is linear. Thus, the change in i at the next timestep (Δv_i^t) is

$$P(\Delta v_i^t | v_i^t, v_j^t, \omega, \sigma, \theta_{ji}) = \omega[\theta_{ji} \cdot v_j^t - v_i^t] + N(0, \sigma) \quad (3)$$

where $\theta_{ji} \in (-\infty, \infty)$ is a multiplier (or “strength”) mapping the value of the cause j to the attractor state of effect i . Figure 2 presents how a variable Y changes as a function of its cause X for a number of different values of θ_{XY} . We assume Δt of 100ms (i.e. between t and $t + 1$) and that ω and σ remain constant, although these assumptions can be loosened (see Lacko, 2012).

The multiple cause case. In general, a variable may have more than one cause. Although there are a variety of ways in which multiple causal influences might combine (cf. Griffiths & Tenenbaum, 2009; Pacer & Griffiths, 2011), here we simply assume that causes have an additive influence on an effects’ attractor state, such that

$$P(\Delta v_i^t | v^t, \omega, \sigma, \Theta) = \omega \left[\left[\sum_j \theta_{ji} \cdot v_j^t \right] - v_i^t \right] + N(0, \sigma) \quad (4)$$

where j now ranges over all causes of variable i and Θ is a square matrix such that $\theta_{ji} \in \Theta$ is the strength of the causal relationship from j to i .² Simply put, the mean that variable i reverts to is assumed to be a sum of the values of its causes, each first multiplied by their respective θ s.

A collection of connected OU processes, which we call an *OU network*, defines causal relationships for all directed relations between variables and unrolls these effects over time. For example, for a system consisting of variables X , Y , and Z , Θ specifies the strengths of the six potential inter-variable causal relationships: $X \rightarrow Y$, $Y \rightarrow X$, $X \rightarrow Z$, $Z \rightarrow X$, $Y \rightarrow Z$, and $Z \rightarrow Y$. Note that non-relationships are specified in this scheme by setting θ_{ji} to zero. At each timestep, Equation 4 is used to determine v_X^{t+1} , v_Y^{t+1} , and v_Z^{t+1} as function of their previous values v_X^t , v_Y^t , and v_Z^t . For display purposes, it is sometimes necessary to constrain v to be between some range. This is done by setting all v^{t+1} that fall outside of the range to their nearest value in the range. The clock then moves forward and the process repeats.

OU networks have some intuitively appealing features of continuously varying causal relationships. Figure 3 demonstrates some of the dynamics that emerge from causal systems simply by varying the θ s. Whereas a positive θ_{XY} results in the value of Y following some positive multiple of the value of X (Figure 3B), a negative θ_{XY} means that a decrease in X drives up the value of Y (e.g. decreasing interest rates is generally thought to increase inflation, Figure 3C). Feedback loops are naturally represented with nonzero values of θ_{XY} and θ_{YX} . A positive feedback loop results if the θ s are of the same sign and have an average magnitude greater than 1 (Figure 3D) whereas a negative feedback loop results if they are less than 1 (Figure 3E). Oscillations can be implemented with θ s of mismatched signs (such as 5 and -5 , Figure 3F). Such feedback loops can be implemented between pairs of variables or as part of a cyclic causal structure with potentially many variables. Combining feedback loops and cycles and including asymmetrical forms can lead to even more complex dynamics (e.g., Figure 3H). We invite the reader to build their own network and observe the dynamics at https://zach-davis.github.io/html/ctcv/demo_ctcv.html.

Inference

We follow Griffiths and Tenenbaum (2005) in modeling people’s learning of causal graphs as inverting the generative model. What must be inferred is the causal structure most likely responsible for producing all variable values at all timepoints— v —under interventions.

Note that to accommodate interventions, we adopt Pearl’s (2009) notion of graph surgery. If variable i is manipulated at time t , the likelihood that v_i^t has its observed value is 1 (i.e., is independent of i ’s previous value or the value of its causes). We define ι_i^t as an indicator variable that is true if variable i is intervened on at t and false otherwise.

The single cause case. Consider the inference problem in which the goal is to determine whether variable j causes variable i and, if so, the sign of that causal relationship. That is, assume a hypothesis space L with three hypotheses. One is that θ_{ji} is greater than 0, a causal relationship we refer to as a *regular connection*. A second is that θ_{ji} is less than 0, referred to as an *inverse connection*. Finally, $\theta_{ji} = 0$ denotes that j has no impact on i .

²Although the OU formalism allows it, throughout this work we ignore the possibility of self-cycles, that is, instances in which variables is a cause of itself. That is, we assume, $\Theta_{ii} = 0$.

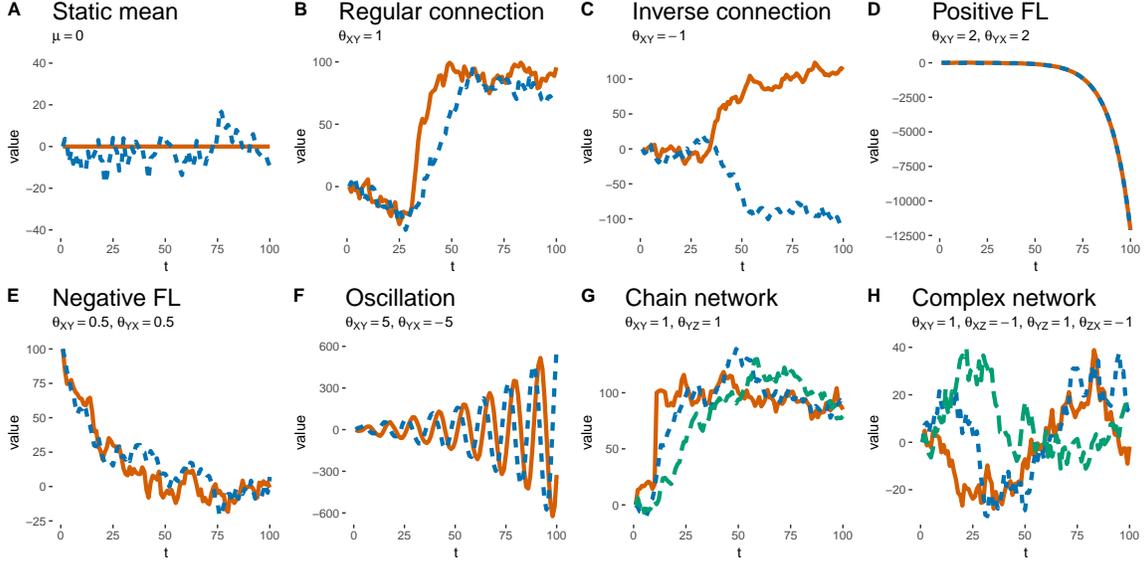


Figure 3. Examples of the dynamical phenomena resultant from varying θ weights. Solid red, dotted blue, and dashed green lines depict the values of variables X , Y , and Z , respectively. **(A)** A system with a single variable Y whose distribution mean is stationary at 0 (i.e., $\mu = 0$). **(B)** A system with variables X and Y and a θ weight from X and Y of 1 (i.e., $\theta_{XY} = 1$). $\mu_X = 0$ for first 30 timepoints and then $\mu_X = 100$ for next 70. The value of Y tracks the value of X . **(C)** The same as Panel B except that X and Y are negatively related ($\theta_{XY} = -1$). The value of Y tracks but has the opposite sign of X . **(D)** A system in which X and Y are reciprocally related via θ weights that are greater than 1 (i.e., $\theta_{XY} = \theta_{YX} = 2$). Because the values of X and Y grow so large they are indistinguishable in the plot. **(E)** The same as Panel D except that X and Y , which have an initial value of 100, are reciprocally related via θ weights that are less than 1 ($\theta_{XY} = \theta_{YX} = .5$). The values of X and Y eventually fluctuate around 0. **(F)** The same as Panel D except that the reciprocal θ s are large and of opposite sign (i.e., $\theta_{XY} = 5, \theta_{YX} = -5$). The values of X and Y oscillate. **(G)** A system with three variables whose θ weights form a causal chain, $\theta_{XY} = \theta_{YZ} = 1$. $\mu_X = 0$ for 10 timepoints but then is set to 100 via an intervention. Note that changes in Y precede changes in Z . **(H)** Timeseries of actual data observed by participant 10 on trial 10, generated by a complex system with three variables and four non-zero θ s. All variables were initialized at 0 and there were no interventions.

Assume that i has no other potential causes.

Computing the posterior probability of a causal hypothesis $l_k \in L$ involves computing, for each timepoint t , the likelihood of the observed change in i (Δv_i^t) given the previous values of i and j (v_i^t and v_j^t), a value of θ_{ji} corresponding to the hypothesis, the endogenous system parameters ω and σ , and any intervention that may have occurred on i (l_i^t). If the learner did not intervene on i at t , this likelihood is given by Equation 3. If they have, it is 1. The product of these likelihoods over all timepoints is proportional to the posterior probability of l_k .

$$P(l_k|v_i, v_j; \iota_i) \propto \prod_t \int_{\omega} \int_{\theta_{ji}} \int_{\sigma} P(\Delta v_i^t | v_i^t, v_j^t, \omega, \sigma, \theta_{ji}; \iota_i^t) P(\theta_{ji} | l_k) P(l_k) P(\omega) P(\sigma) d\sigma d\theta_{ji} d\omega \quad (5)$$

$P(\omega)$ and $P(\sigma)$ represents the learner’s prior beliefs about ω and σ . $P(\theta_{ji} | l_k)$ represents the priors over θ_{ji} corresponding to hypothesis l_k . For example, if l_k corresponds to a regular connection, $P(\theta_{ji} | l_k)$ would be 0 for non-positive values of θ_{ji} . For positive values, it would reflect learner’s priors over θ_{ji} for regular connections. (Later we describe how these priors can be estimated in our experiment on the basis of an instructional phase that precedes the causal learning task.) Applying Equation 5 to each causal hypothesis and then normalizing yields the posterior over the three hypotheses in L .

A complication arises if variable values v are truncated between some range of values (in our task $v \in [-100, 100]$). In the case where v_i^t equals the maximum truncated value, the likelihood is the mass of the likelihood distribution above the range of values. For the minimum truncated value the likelihood is the mass of the likelihood distribution below the range of values.

The multiple cause case. This procedure for evaluating a single potential causal relationship generalizes to determining the structure of an entire OU network. Consider a hypothesis space G as consisting of *graphs* where each graph defines, for every potential causal relationship, whether it is positive, inverse, or zero. For a system with n variables G would contain 3^{2n} distinct causal hypotheses; for our example system with variables X , Y , and Z , G contains 729 graphs. The posterior probability of a graph $g_k \in G$ involves computing for each variable i and timepoint t , the likelihood of the observed Δv_i^t given the θ s defined by g_k and the state of the system’s variables at t (Equation 4), taking into account the possibility of an intervention on i at t (ι_i^t):

$$P(g_k | v; \iota) \propto \prod_{i=1}^N \prod_t \int_{\omega} \int_{\theta} \int_{\sigma} P(\Delta v_i^t | v^t, \omega, \sigma, \theta; \iota_i^t) P(\theta | g_k) P(g_k) P(\omega) P(\sigma) d\sigma d\theta d\omega \quad (6)$$

Experiment: Causal Structure Learning

To test people’s ability to learn causal structure between continuous variables in continuous time, we conducted an experiment in which participants freely interact with sliders governed by an OU network with hidden causal structure. Their goal was to intervene on the system in order to discover the hidden causal structure.

Method

Participants. 30 participants (13 female, age $M = 37.5$, $SD = 10.6$) were recruited from Amazon Mechanical Turk using psiTurk (Crump, McDonnell, & Gureckis, 2013; Gureckis et al., 2016). They were paid \$4 for approximately 30 minutes. In a post test questionnaire, on a ten point scale participants found the task engaging ($M = 7.9$, $SD = 2.2$) and not particularly difficult ($M = 3.9$, $SD = 2.6$).

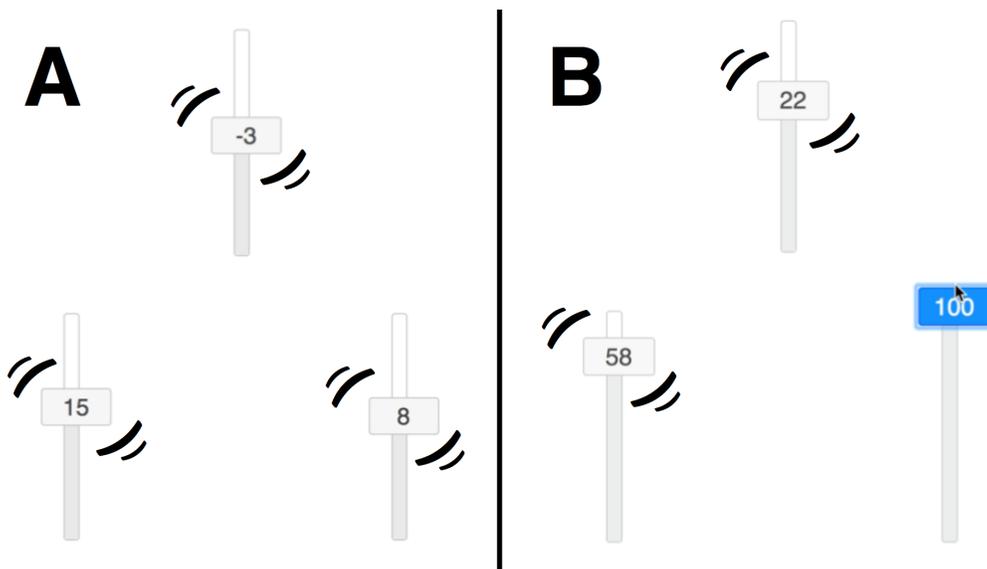


Figure 4. Sliders used by participants. (A) shows that the sliders all jitter if no interventions are made. (B) shows that the sliders do not jitter if intervened on.

Materials. Each of the three variables was represented by a vertical slider that moved by itself according to the underlying OU network but which could also be manipulated by clicking and dragging anywhere on the slider, overriding the state it would otherwise have taken (see Figure 4).³ A timer was presented at the top of the screen. Participants responded using six additional sliders presented beneath the trial window, one for each potential causal relations. Responses were constrained to be one of three options: ‘Inverted’, ‘None’, or ‘Regular’, corresponding to $\theta < 0$, no relationship ($\theta = 0$), and $\theta > 0$, respectively. Participants were pretrained on these terms in the instructions. The sliders were constrained to be between -100 and 100, and the buttons on the slider presented a rounded integer value in addition to moving up and down.

Stimuli and Design. The 23 causal graphs shown in Figure 5 were selected for testing on the basis of a number of criteria. They were roughly balanced in the number of positive and negative links and the number of links between each of the variables. More qualitatively, we tried to select networks that would be interesting a priori. This includes many of the classic causal graphs such as chain networks, common causes, and common effects, but also less-studied graphs such as those with feedback loops. The experiment always began with two practice trials that were excluded from all analyses. These were always the two *Single cause* networks (Figure 5, top left). This was followed by 23 test trials, one for each of the networks in Figure 5 presented in random order. The OU parameters used during training and the test were $\omega = .1$ and $\sigma = 5$. The true θ s were either 1 (for regular connections), 0 (no connection), or -1 (for inverse connections).

Procedure. To familiarize them with the interface, participants were required to first watch four videos of an agent interacting with example causal networks. These videos informed participants of the underlying causal structure and demonstrated an agent inter-

³See <https://zach-davis.github.io/publication/cvct/> for a demo.

Type	Network	Type	Network
Single cause		Feedback loop	
Chain network		Feedback loop with feed-out	
Common cause		Feedback loop with feed-in	
Common effect		Feedback loop with chain	
Simpson's paradox			

Figure 5. All 23 structures participants were tasked with learning. Black arrowheads signify “regular” connections ($\theta = 1$), white arrowheads signify “inverse” connections ($\theta = -1$).

acting with the system. To minimize biasing participants toward any particular intervention strategy, the videos displayed a variety of basic movements, including wobbling the intervened on variable, holding a variable at a constant level, and holding a variable at a limit value (e.g., 100) by moving its slider to one end of the scale. The four example causal networks included (1) no causal connections, (2) a single regular ($\theta = 1$) connection, (3) a single inverse ($\theta = -1$) connection, and (4) two connections forming a causal chain in which one link was regular and one was inverse. To ensure that they understood the task, participants were required to pass a five question comprehension check before starting. If a participant responded incorrectly to any of the five questions they were permitted to retake the quiz until they responded correctly to all five questions. This was designed to ensure that they learned: the duration of each trial, the difference between a regular and inverted connection, that there can be more than one connection per network, and that they must provide a response for all six possible connections.

In the main task that followed, participants initiated a trial by pressing the “Start” button. The sliders started moving, with the values updating every 100ms. Perceptually, they would appear to “jitter” according to the noise associated with the underlying OU network plus move systematically according to the unknown causal relationships. At any time, participants were free to intervene on any variable by clicking, holding, or dragging the requisite slider. While it was pressed down, the position of the mouse determined the value of the variable. Once it was released the variable would continue from that point according to the OU network. Participants were free to make (and revise) their judgments at any point after initiating a trial but were required to enter a judgement for all six causal relations by the end of the trial (see Figure 6). No feedback was provided at any point. After completing the 25 trials, participants completed a brief post-test questionnaire reporting their age, gender, engagement and subjective difficulty as well as any comments.

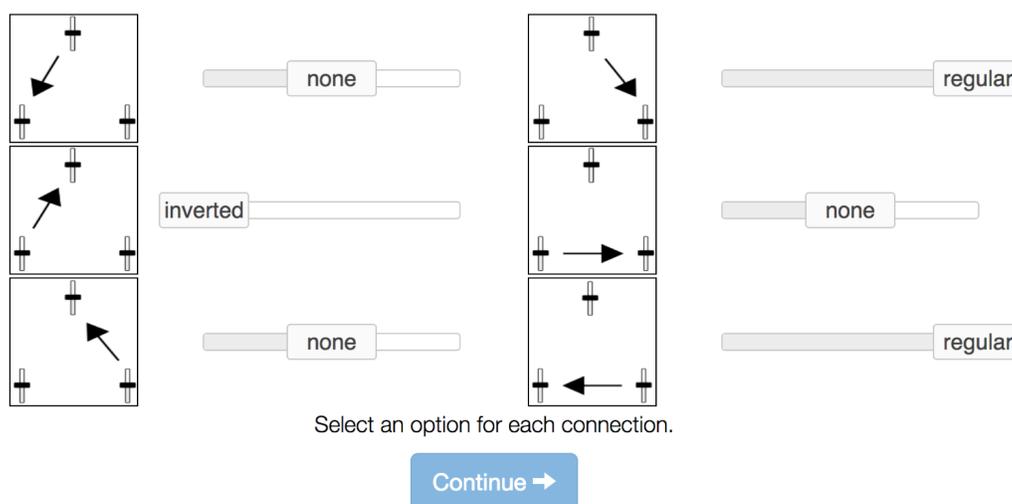


Figure 6. Judgment options for participants. Participants were presented with a ternary choice between “inverted”, “none”, and “regular”.

Results

Participants were substantially above chance (.33) in correctly classifying causal links into one of the three response categories ($M = .82$, $SD = .22$), $t(29) = 17.48$, $p < .001$. They were slightly more successful in identifying regular causal links ($M = .92$, $SD = .12$) than inverse causal links ($M = .90$, $SD = .13$), $t(29) = 2.12$, $p = .04$. Participants also correctly classified a higher proportion of causal relationships as the trials progressed, as demonstrated by a simple linear regression of accuracy on trial number, $t(21) = 2.91$, $p = .008$, although this relationship was modest with participants being 0.25% more likely to correctly identify a link for each new trial.

In identifying overall causal networks (correctly identifying all 6 of the possible directional causal relationships), participants were also well above chance ($3^{-6} = .0014$), ($M = .44$, $SD = .22$), $t(29) = 10.81$, $p < .001$. The probability of selecting the correct network was .79, .60, .25, and .07 for networks with 1, 2, 3, and 4 causal links, respectively. Accuracy varied sharply with the complexity of model as shown by a repeated measures ANOVA, $F(3, 84) = 74.0$, $p < .001$. Note that participants’ responses did not reflect a preference toward simpler models, as they marked slightly over half of the possible connections ($M = .52$, $SD = .13$), which was greater than the true proportion of connections in the test networks (.39), $t(29) = 5.62$, $p < .001$. See the Supplementary Materials for results for all tested networks.

Errors. While participants were generally well above chance in identifying causal relationships, there was some systematicity to their errors. In particular, these errors closely followed the qualitative predictions of Fernbach and Sloman’s (2009) local computations (LC) model. The first qualitative prediction is an over-abundance of causal links. Eighty-two percent ($SD = .17$) of the errors that participants made involved adding extraneous

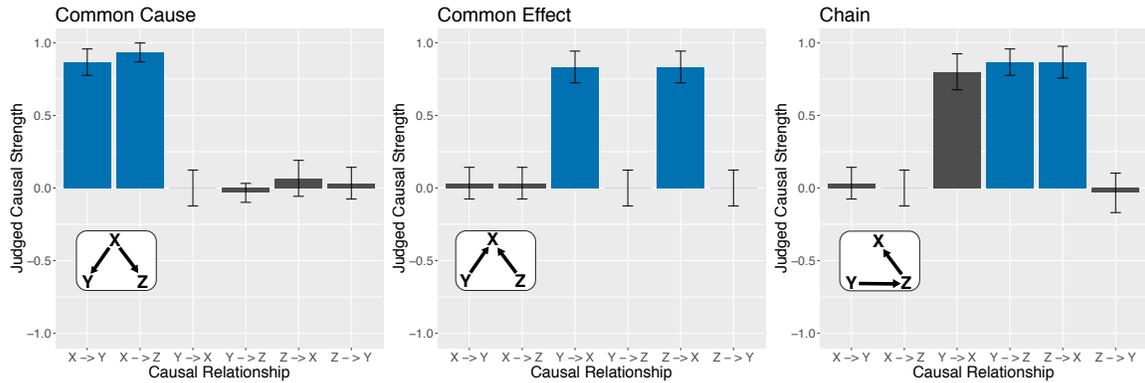


Figure 7. Participant judgments of causal relationships for three tested networks. Bar colors correspond to the true causal structure, namely, blue for regular connections and gray for no connection. Bar heights represent mean θ reported by participants (regular = 1 and none = 0). Because these networks included only regular causal relationships, no instances of inverse relationships are shown. Error bars denote 95% confidence intervals.

causal links, significantly greater than chance⁴ (.59); $t(29) = 7.33$, $p < .001$. The second qualitative prediction of the LC model as defined in this paper is an inability to distinguish between direct and indirect causes (e.g. in the network $X \rightarrow Y \rightarrow Z$, incorrectly also judging $X \rightarrow Z$). While in general participants correctly classified 82% of the causal links, they were far more likely to erroneously add a direct link between two variables when in fact the relationship between those variables was mediated by a third variable ($M = .16$, $SD = .21$), performing below chance (.33); $t(29) = -4.48$, $p < .001$.

Figure 7 shows participant judgments for three classic causal structures in causal cognition: common cause, common effect, and chain networks. It shows that participants were quite good at detecting any causal relationship in a network that existed between two variables. In the figure, these results correspond to the blue bars, which indicate that they correctly classified a regular connection as regular. (As mentioned, participants were also good as classifying inverse connections as inverse.) Figure 7 also shows that participants were often good at classifying absent connections as absent (the gray bars) with one important exception: in the the chain network $Y \rightarrow Z \rightarrow X$ the relationship between Y and X was judged to be nearly as causal as $Y \rightarrow Z$ and $Z \rightarrow X$. That is, they failed to appreciate that the (apparent) relationship between Y and X was in fact mediated by Z . These patterns held for the other instances of the common cause, common effect, and chain networks defined in Figure 5. Moreover, we found that, for any of the more complex networks in Figure 5, participants had a strong tendency to infer a direct causal relationship between two variables whenever those variables were in fact mediated by the third variable. Figure 8 presents how causal links were classified for all 23 networks.

Interventions. To achieve this level of performance, participants made heavy use of interventions. We define a single intervention as beginning when a participant clicked on a variable’s slider and ending when the mouse was released. The average number of inter-

⁴For the structures used in this experiment, a hypothetical participant who responded “inverse”, “none”, and “positive” with equal probability would erroneously add a causal link 59% of the time.

ventions made on a single trial was 4.94 (SD = 2.46). However, because a few participants made a large number of interventions on most trials, this distribution was modestly skewed with a median of 4 and mode of 3. One participant made no interventions at all.

Interventions lasted an average of 3.46 seconds (SD = 3.00) and had a range (the maximum value of the variable during the intervention subtracted from its minimum value) of 138.3 (SD = 58.89). This latter measure was strongly bimodal with modes around 100 and 200, indicating that interventions typically consisted of participants dragging a variable from about 0 to one end of the scale (-100 or 100) or then in addition dragging it to the opposite end of the scale. Apart from these large swings, participants typically held the variable steady at a constant value during an intervention. This conclusion is supported by the fact that, within an intervention, the percentage of 100ms time windows in which the variable had the same value as during the previous window was 71.2%. Four participants had some tendency to “wobble” the variable through a small range during an intervention but they were the exception.

The interventions were spread about evenly over the three variables. Indeed, all three network variables were manipulated at least once on more than 99% of the trials. Interventions varied modestly as a function of whether the manipulated variable was a cause of other variables in the network. When it was, the intervention was both shorter (3.21s) and had a narrower range (132.9) than when it wasn’t (3.99s and 149.5), $t(28) = 3.19$ and $t(28) = 6.39$, respectively, both $ps < .005$.⁵ Apparently, it was easier for participants to identify causes, which involves observing a state change in other network variables, than non-causes, which involves the absence of such changes. Interventions on causes did not vary substantially, in length of time or range of values, as a function of whether they had one or two effects. Interventions also did not vary as a function of whether or not the variable was affected by other variables in the network. In summary, participants recognized that interventions help causal learning, that manipulating all variables is necessary to identify the correct causal structure, and that large interventions are more useful than small ones.

Results summary. Participants exhibited considerable ability to intervene effectively and learn causal structure in our task. Despite these abilities, they also made systematic errors consistent with the predictions of the LC model. It is not clear whether the data considered as a whole is more consistent with normativity or a more locally focused model. Indeed, it is not even clear that participants are using the OU functional form to infer connections, rather than a more general model such as one that assumes linearity. For a more granular analysis of people’s causal structure learning, we now turn to a number of theoretical accounts of how people learn causal structure.

Modeling

In this task we compare a total of nine models corresponding to different accounts of how people learn causal structure. These accounts can be roughly categorized as modeling people as normative, local, linear, or random in their causal learning behavior. We compare the ability of these models’ to predict participants’ causal structure judgments.

⁵There were 28 degrees of freedom for these analyses, rather than 29, because one of the 30 participants did not intervene.

OU models

Normative model. Normative inference for the current task requires that a learner maintain a distributional belief over all possible causal structures and update it according to the data they experience. Equation 6 above defines normative inference in this task. There has been much work suggesting that adults and children are capable learners of causal structures and act roughly in accordance with the normative model, at least in sufficiently simple scenarios (Gopnik et al., 2004; Griffiths & Tenenbaum, 2009). We ask whether these conclusions generalize to the sort of causal systems under investigation here.

Recall that Equation 6 assumes that learners have priors over ω , σ , and the θ s. We assume for simplicity that learners acquire a rough approximation of the true values of these parameters (i.e., $\omega = .1$, $\sigma = 5$, and $\theta \in (-1, 0, 1)$) while watching the four instructional videos, but assume some spread to accommodate uncertainty. The distributions we assumed over parameters were thus⁶,

$$\theta \sim \Gamma(\text{shape} = 5 \times \theta_{true}, \text{rate} = 5)$$

$$\omega \sim \Gamma(\text{shape} = 100 \times \omega_{true}, \text{rate} = 100)$$

$$\sigma \sim \Gamma(\text{shape} = 100 \times \sigma_{true}, \text{rate} = 100)$$

Note that θ values are defined by the graph. For regular connections, θ is distributed as above. For inverse connections, the sampled values are negated. For non connections θ is 0.

Local computations model. We compare the normative model to a “local computations” (LC) model that has been advocated as a general-purpose account of causal learning behavior (Bramley, Dayan, et al., 2017; Fernbach & Sloman, 2009). Applied to an OU network, the LC model entails deciding, for each potential causal relationship considered in isolation, whether the observed values of those two variables implies a regular, inverted, or zero causal relation. It thus involves applying Equation 5 above to each potential causal relationship. The LC model assumes the same priors over ω , σ , and the θ s as the normative model.

A key distinction between the normative and LC models of course is their ability to detect whether a relationship between two variables is mediated by a third. For example, in the network $X \rightarrow Y \rightarrow Z$, X and Z have many of the hallmarks of a direct causal relationship: They are correlated, changes in X precede changes in Z , and intervening on X later affects Z (but not vice versa). Whereas the normative model would take into account the mediated relationship between X and Z (by noting the absence of an X/Z correlation when controlling for Y), LC, which evaluates individual causal links without consideration of the entire graph, would not recognize the mediating role of Y and so infer $X \rightarrow Z$ in addition to $X \rightarrow Y$ and $Y \rightarrow Z$. Of course, we have already seen partial evidence that participants may be poor at detecting mediated relationships (Figure 7). Modeling will reveal whether the LC model is a good account of all the data, or if it only accounts for participants’ errors.

⁶The 10% and 90% quantiles associated with these distributions are .49 and 1.60 for θ , .062 and .142 for ω , and 4.72 and 5.29 for σ .

Alternative models

We compare the two OU-based models to alternatives that assume linear relationships between cause and effect. In particular, we compare two approaches to modeling timeseries information from the literature: time-lagged correlation and Granger causality. Each of these approaches is applied to three candidate representations for learning causal structure between continuous variables, as introduced by Soo and Rottman (2018); *state representations*, *difference scores*, and *trinarized difference scores*.

In these linear models, the value of variable i at time t is modeled as

$$P(v_i^t | v^{t-1}, \sigma, \beta) = \sum_j [\beta_{ji} \cdot v_j^{t-1}] + N(0, \sigma) \quad (7)$$

where j denotes all causes of variable i (including i itself) and β_{ji} denotes the partial slope coefficient or strength of that cause on the effect. Analogously to our treatment of θ values in the OU models, for the linear models we assume some uncertainty about the strength parameter $p(\beta)$ but that these differ in sign for regular and inverse connections, and also model people as having uncertainty over standard deviation $p(\sigma)$. The marginal likelihood of v_i for a graph thus involves computing, for each timepoint, the likelihood of that variable’s value given the β predictors defined by the graph and the value(s) of its cause(s), and marginalizing over $p(\beta)$ and $p(\sigma)$. We treat interventions in the same manner as the OU models. As before, we compute the total likelihood as the product of the marginal likelihoods of all variables at all timepoints under each graph, assume an initially uniform prior over graphs and compute the resulting posterior. The unnormalized posterior probability of a causal graph given all values of all variables at all timepoints is thus

$$P(g_k | v; \iota) \propto \prod_t \prod_i \int_{\beta} \int_{\sigma} P(v_i^t | v^{t-1}, \sigma, \beta; \iota_i) P(\beta | g_k) P(g_k) P(\sigma) d\sigma d\beta \quad (8)$$

This general procedure can be applied to each of the linear models by modifying the state representation v or prior over β . For the three candidate representations introduced by Soo and Rottman (2018): State representations involves inference over the actual variable values; difference scores involves inference over variable values after computing $v^t - v^{t-1}$; trinarized difference scores involves inference over difference scores that have been converted to -1 when negative and 1 when positive.

The difference between time-lagged correlation and Granger causality is just whether β_{ii} is included as a predictor, that is, whether v_i^t is influenced by v_i^{t-1} as well as its causes. Granger causality includes this term while Time-lagged correlation does not.

Unlike the OU models, there is no natural ground truth parametrization for the linear models on which to center reasonable distributional parameter beliefs. Thus we must find another way to choose reasonable settings for $p(\beta)$ and $p(\sigma)$. We chose the mean of our distributions by fitting the $\hat{\beta}_{ii}$, $\hat{\beta}_{ji}$, and $\hat{\sigma}$ values that maximized the posterior probability of the true causal graphs across all subject data (including β_{ii} for the Granger models). We then made analogous assumptions about the spread around these means as we did for θ and σ in the OU models—namely,

$$\beta \sim \Gamma(\text{shape} = 5 \times \hat{\beta}, \text{rate} = 5)$$

Table 1
Summary of Model Accuracy and Performance

Model	State representation	Accuracy	Judge	BIC	Px
OU local computations		.89	.82	6,163	21
OU normative		1.00	.82	6,475	4
Granger causality	states	.91	.78	7,079	1
	difference scores	.82	.69	8,415	1
	trinarized diff scores	.49	.42	9,859	0
Time-lagged correlation	states	.89	.74	7,901	1
	difference scores	.82	.69	8,407	0
	trinarized diff scores	.63	.50	9,793	0
Baseline		.17	.17	9,888	2

Note: *Accuracy* = proportion of links drawn that match ground truth. *Judge* = proportion of links drawn that match participant judgments *BIC* = Bayesian Information Criterion *Px* = number of participants best fit by that model

$$\sigma \sim \Gamma(\text{shape} = 100 \times \hat{\sigma}, \text{rate} = 100).$$

β values are treated the same as in the OU models. Regular connections are distributed as above, inverse connections are negated.

Comparing the models

We compare participants' structure judgments to the predictions of these models across all the test trials in our experiment. In total, we consider nine models. These are eight described above: (1) *normative*, (2) *local computations (LC)*, and three variants of both (3-5) *Granger causality* and (6-8) *Time lagged correlation* varying whether they were based directly on states, difference scores, or trinarized difference scores. Finally, we compare these against (9) a *Baseline* model that assumes each judgment is a random selection from the space of possible graphs. We marginalized over θ , ω , σ by drawing 1,000 samples from their respective distributions and averaging the likelihood within each causal model. To account for decision noise in selecting causal graphs from their posterior distributions, for each model apart from the baseline we fit (by maximum likelihood using R's `optim` function) a single softmax parameter τ that maximized the posterior probability of participant selections.

Results and discussion. Table 1 details the results of our comparison. For each inference model we report the overall proportion of the true connections identified across all trials assuming the most probable graph is selected at the end of each trial (Accuracy column), the proportion of participant's edge judgments that correspond with the most probable graph under the model (Judge column), the Bayesian Information Criterion of all participant's judgments according to that model (BIC column); and the number of participants best fit by each model.

Unsurprisingly, the normative model was the most successful at recovering the underlying structure, but many other models were also successful. The only models that struggled were those that used trinarized difference scores as their representation, showing that the magnitude of changes in the variables is important to capturing the structure of the data.

Next, we compared the MAP estimates of causal structure of the models to participant judgments. In this coarse measure, the OU models were roughly equal to each other in matching participant judgments, and were also similar to some of the linear models.

The results of the more sensitive posterior probability analysis were clearer in distinguishing between models. Over all participants, the LC model had the highest log-likelihood. On a per participant basis, of the 30 participants 21 were best fit by the LC model, with the normative model being the best account of 4 participants. The remaining 5 participants were split among the linear models or were at baseline.

General Discussion

In this paper, we introduced a generative model of causal influence relating continuous variables over time. We showed how such systems can exhibit emergent behaviors such as excitatory or inhibitory feedback and oscillations, depending on specific settings of relative causal strengths between variables. When learning from this rich data, people were best described as considering individual pairs of variables, rather than updating their beliefs over entire structures. This finding accords with an intuitive description of how people handle continuous information flowing in real time: they focus their attention on smaller, more manageable problems rather than attempting to tackle the full torrent of information.

Local inference

A key result in our task was that most participants evaluated pairwise relationships between variables rather than updating their beliefs over all possible causal structures. This conclusion was drawn from the superior fit of the locally focused LC model, and corroborated by qualitative results such as the finding that participants often inferred direct causal relationships between variables that were in fact only indirectly related (through a third mediating variable). These results are consistent with previous findings suggesting that, rather than representing a full hypothesis space, people tend to consider a single hypotheses to which they make small alterations (Bramley, Dayan, et al., 2017; Fernbach & Sloman, 2009; Quine, 1969). Here we show that this principle of causal learning extends to much richer scenarios. Indeed, it may be the case that real time continuous information places stronger demands on attention and memory than the original settings that provided evidence for the LC model. If this were true, it would be especially reasonable to use the resource-efficient local strategy in these more demanding environments.

A potential alternative conceptualization of the LC model is that it instantiates the idea that distal causes are still considered as causal. For example, most people would not find it inappropriate to say that the reintroduction of wolves to Yellowstone National Park caused changes to the ecosystem, even if many of these changes came indirectly through other variables such as changes in the movement of elk (Fortin et al., 2005). While this is a reasonable conceptualization, we believe that it is not as good an account of our data as the LC model. For one, we explicitly provided participants with an example in the instructions that showed the movement of a chain network without the additional indirect connection. This should have reduced the possibility that participants were unclear about whether they should consider distal causes as causal. This accords with findings in the literature that people exhibit locality despite feedback, incentives, and explicit instruction with examples

that encourage people to not draw the additional causal link (Bramley, Dayan, et al., 2017; Bramley, Lagnado, & Speekenbrink, 2015; Fernbach & Sloman, 2009). More fundamentally, this “distal” account makes assumptions about how people are approaching the task that we consider unlikely. It models them as doing full normative inference, and then having a response bias to draw indirect connections. Figure 8 shows that indirect connections were less likely to be responded to as causal than the direct connections, which would imply a response bias where participants have the full causal model but would only on occasion draw the additional indirect connection. The LC model, in contrast, naturally considers indirect connections as less causal due to the underlying dynamics of OU networks. While indirect causal relationships do have many hallmarks of direct causal relationships (correlation, temporal asymmetry, asymmetric results of interventions), they are not identical. In $X \rightarrow Y \rightarrow Z$, changes to Z in response to X are more temporally removed and noisier than would be predicted if there were a direct $X \rightarrow Z$ connection, and therefore the LC model assigns a lower (but still reliably non-zero) probability to these potential connections. Because the LC model accounts for the patterns of errors as naturally arising from the interaction of system dynamics and cognitive limitations, rather than as a response bias over normative inference, we consider it a better account of the behavior of participants in our task.

Interventions

One contribution of the OU network framework is the introduction of a qualitatively different type of intervention. In a typical study of causal cognition learners are able to, on a particular trial, turn a variable on or off and observe the values of other variables. In contrast, interventions in our task are extended through time and can encompass a wide range of variable values. Participants generally recognized that the most informative actions involved large swings in variable values and systematic manipulation of each variable in the system.

Nevertheless, note that while their interventions were informative they were less than optimal. In fact, the most efficient interventions in this task involve rapid swings between the ends of the variable’s range. But whereas participants used the full range, they tended to hold a variable at one value for longer than necessary. Doing so yields useful but somewhat redundant information. Of course, perhaps this strategy reflected participants’ need for redundant information imposed by cognitive processing limits. It may also reflect their inability or unwillingness to engage in the rapid motor movements required by the optimal strategy.

Although participants could intervene on any variable at any time to set it to any value, they were constrained to manipulating one variable at a time. Future studies could expand the action space by, for example, allowing participants to “freeze” one variable at a value while manipulating others. Of course, an ability to “control for” one variable while investigating the relationship between two others might help learners identify mediating relationships. For example, freezing Y and then manipulating X in $X \rightarrow Y \rightarrow Z$ would result in no change in Z , perhaps reducing the chance that the learner would conclude $X \rightarrow Z$.

Future directions

The proposed OU network framework can be extended across a variety of dimensions in future research. For example, in this paper’s instantiation of OU networks, a cause impacts an effect on the next timepoint. The impact of a cause on effect could be distributed over multiple timepoints, or at some stochastically selected timepoint. Such studies could contribute to debates about the influence of time on causal learning, such as that judgments of causality are strengthened by temporal contiguity (Shanks et al., 1989) or the reliability of delays (Bramley et al., 2018; Buehner & May, 2003). Varying the gap between timepoints (in this task t to $t+1$ was 100ms) may result in different approaches by participants. Use of continuous variables naturally allows consideration of a greater number functional forms relating causes and effects (Griffiths & Tenenbaum, 2009). Latent causes can be introduced to model implicit inference of mechanisms relating cause and effect. Complex, non-linear data can be generated to study people’s learning from time series data (Caddick & Rottman, 2019; Soo & Rottman, 2018). The outcomes of experiments using these richer causal systems will help to evaluate the generalizability of models of causal cognition that have heretofore been tested mostly on Bayes nets applied to discrete events.

The formalism developed in this paper also has potential application to the domain of control. Many aspects of everyday life, as well as interesting domains in AI and machine learning, can be classed as control problems in which there is initial or ongoing uncertainty about the structure of the control domain. Some studies have explored settings where participants attempt to manipulate a simple environment that is reactive to their decisions to maximise gain (for review, see Osman, 2010). However, one limitation of extant work is that the environments typically used are ad hoc and no class of environments has been thoroughly explored. In parallel, much recent attention in machine learning has been given to demonstrations of successful control in small worlds such as atari and board games. However, generalisation to new goals or related environments continues to be poor (Lake, Ullman, Tenenbaum, & Gershman, 2017). In recent work, we propose OU networks as a systematic class of control environments. This approach allows research into human control to ask new questions, such as what structures are inherently easy or hard to identify or control and under what circumstances does successful control depend on an accurate model of a system’s structure (Davis, Bramley, Rehder, & Gureckis, 2018).

Functional form

Given people’s well-known bias towards assuming linear functional forms (Brehmer, 1974; Byun, 1996; DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Griffiths, & Lewandowsky, 2007; Kalish, Lewandowsky, & Kruschke, 2004; Kwantes & Neal, 2006), it may be a surprising result that the alternative models assuming linearity did not match people’s judgments as well as those using the Ornstein–Uhlenbeck functional form. This result has a number of possible explanations. For one, as discussed before, Ornstein–Uhlenbeck processes appear to be relatively common across a range of domains, and people may have a developed representation of the functional form that they brought to the task. It is also possible that participants do not have a direct representation of Ornstein–Uhlenbeck processes, but were able to recognize higher-order movement statistics that are not present in linear models (e.g. OU processes, unlike linear relationships, exhibit acceleration toward their attractor

basin). For example, people may have applied a general function approximator such as a Gaussian Process to the relationship between cause and effect and abstracted a function closer to OU processes than linearity. Future work could explore settings where learning the functional form between cause and effect is not possible (such as one-shot learning) or settings where the impact a cause has on its effect is linear.

Limitations

There are a number of limitations to the current project that could be addressed with further experiments. For one, while we did account for uncertainty over parameters of our models, we did not account for other sources of noise such as the likelihood that people cannot attend to all three variables simultaneously.⁷ This issue will likely compound as more variables are added. Additionally, the presented analyses in this paper discuss but do not model intervention decision-making, a critical component of the active learning of causal structure. Future analyses would naturally involve, as a benchmark to compare against humans, models for selecting actions that maximize expected information gain.

Conclusions

We have no doubt that the canonical causal relationships between discrete events (e.g., take a pill \rightarrow headache relieved) that have been the main focus of causal cognition often serve as highly useful and approximately correct parts of human’s semantic representation of the world. But sometimes details matter. Causal influences emerge over time, may reflect functional relationships that are as complex as the underlying mechanisms that produce them, and afford interventions that vary in their duration and intensity. Complex patterns of feedback may be the rule rather than the exception (Cartwright, 2004; Sloman & Lagnado, 2015; Strevens, 2013). Apprehending these properties may even be a precondition to forming the (highly summarized and approximate) causal relations between discrete events that are so simple to represent and easy to communicate.

We instantiated a learning task in which people were confronted with some of these challenges, including continuously-observed continuous variables, feedback cycles, and the ability to carry out extended interventions. We found that they exhibited considerable success identifying the correct causal structure but also committed systematic errors, errors consistent with a model that describes people as narrowly investigating individual causal relationships rather than updating their beliefs wholesale. We hope that the formalism presented in this paper will help spur greater study of the mechanisms for learning and action in this important class of problems.

References

- Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.
- Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

⁷Although Vul, Alvarez, Tenenbaum, and Black’s (2009) finding that people optimally allocate attention to particles moving according to an OU process may ameliorate the latter concern

- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.
- Bramley, N. R., Gerstenberg, T., & Lagnado, D. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).
- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880.
- Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through interventions. *Journal of Experimental Psychology: Learning, Memory & 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 *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.
- Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, *56*(5), 865–890.
- Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning*, *15*(2), 167–196.
- Byun, E. (1996). *Interaction between prior knowledge and type of nonlinear relationship on function learning*. (Unpublished doctoral dissertation). ProQuest Information & Learning.
- Caddick, Z. A., & Rottman, B. M. (2019). Politically motivated causal evaluations of economic performance.
- Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, *71*(5), 805–819.
- Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367.
- Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, *79*, 102–133.
- Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.
- Davis, Z., Bramley, N., Rehder, B., & Gureckis, T. M. (2018). A causal model approach to dynamic control. In *Proceedings of the 40th annual conference of the cognitive science society*. Madison, WI: Cognitive Science Society.
- DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.
- Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1327.

- Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, 35(3), 678.
- Fortin, D., Beyer, H. L., Boyce, M. S., Smith, D. W., Duchesne, T., & Mao, J. S. (2005). Wolves influence elk movements: behavior shapes a trophic cascade in yellowstone national park. *Ecology*, 86(5), 1320–1330.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1), 3.
- Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, 139(4), 756.
- Griffiths, T. L. (2004). *Causes, coincidences, and theories* (Unpublished doctoral dissertation). Stanford University.
- 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.
- Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, 48(3), 829–842.
- Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, 133(3), 611–620.
- Hitchcock, C. (1997). Probabilistic causation.
- Hume, D. (1740). An enquiry concerning human understanding. In *Seven masterpieces of philosophy* (pp. 191–284). Routledge.
- Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2), 288–294.
- Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, 111(4), 1072.
- Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430.
- Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1019.
- Lacko, V. (2012). Planning of experiments for a nonautonomous ornstein-uhlenbeck process. *Tatra Mountains Mathematical Publications*, 51(1), 101–113.
- Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 451.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological review*, 115(4), 955.
- Marsh, J. K., & Ahn, W.-k. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning*,

- Memory, and Cognition*, 35(2), 334.
- McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(2), 395.
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychological bulletin*, 136(1), 65.
- Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. In *Proceedings of the 24th international conference on neural information processing systems* (pp. 2384–2392).
- Pacer, M. D., & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Quine, W. V. O. (1969). *Word and object*. MIT press.
- Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, 72, 54–107.
- Rothe, A., Devereitt, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, 179, 266–297.
- Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: observations and interventions. *Cognitive psychology*, 64(1), 93–125.
- 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.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.
- Sloman, S. A., & Lagnado, D. (2005). Do we “do”. *Cognitive Science*, 29, 5–39.
- Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, 66, 223–247.
- Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, 147(4), 485.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, 27(3), 453–489.
- Strevens, M. (2013). Causality reunified. *Erkenntnis*, 78(2), 299–320.
- Taylor, E. G., & Ahn, W.-k. (2012). Causal imprinting in causal structure learning. *Cognitive psychology*, 65(3), 381–413.
- Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, 36(5), 823.
- Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in neural information processing systems* (pp. 1955–1963).
- Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 216.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology:*

General, 121(2), 222–236.

Supplement

Figure 8 presents results for all 23 tested networks. Note that although overall performance was good, participants consistently reported the presence of a direct causal relationship between variables whenever that relationship was in fact mediated by the third variable. For example, in the first instance of a Chain network in the figure ($Y \rightarrow Z \rightarrow X$), participants incorrectly judged that $Y \rightarrow X$ (i.e., ignored the mediator Z). In the first instance of a Feedback Loop W/Feedout ($Z \leftrightarrow Y \rightarrow X$), they judged that $Z \rightarrow X$ (ignoring mediator Y). In the first instance of a Feedback Loop W/Feedin ($X \rightarrow Z \leftrightarrow Y$), they judged that $X \rightarrow Y$ (ignoring mediator Z). In the first instance of a Feedback Loop W/Chain ($Y \rightarrow Z \leftrightarrow X \rightarrow Y$), they judged that $Y \rightarrow X$ (ignoring mediator Z) and $Z \rightarrow Y$ (ignoring mediator X).

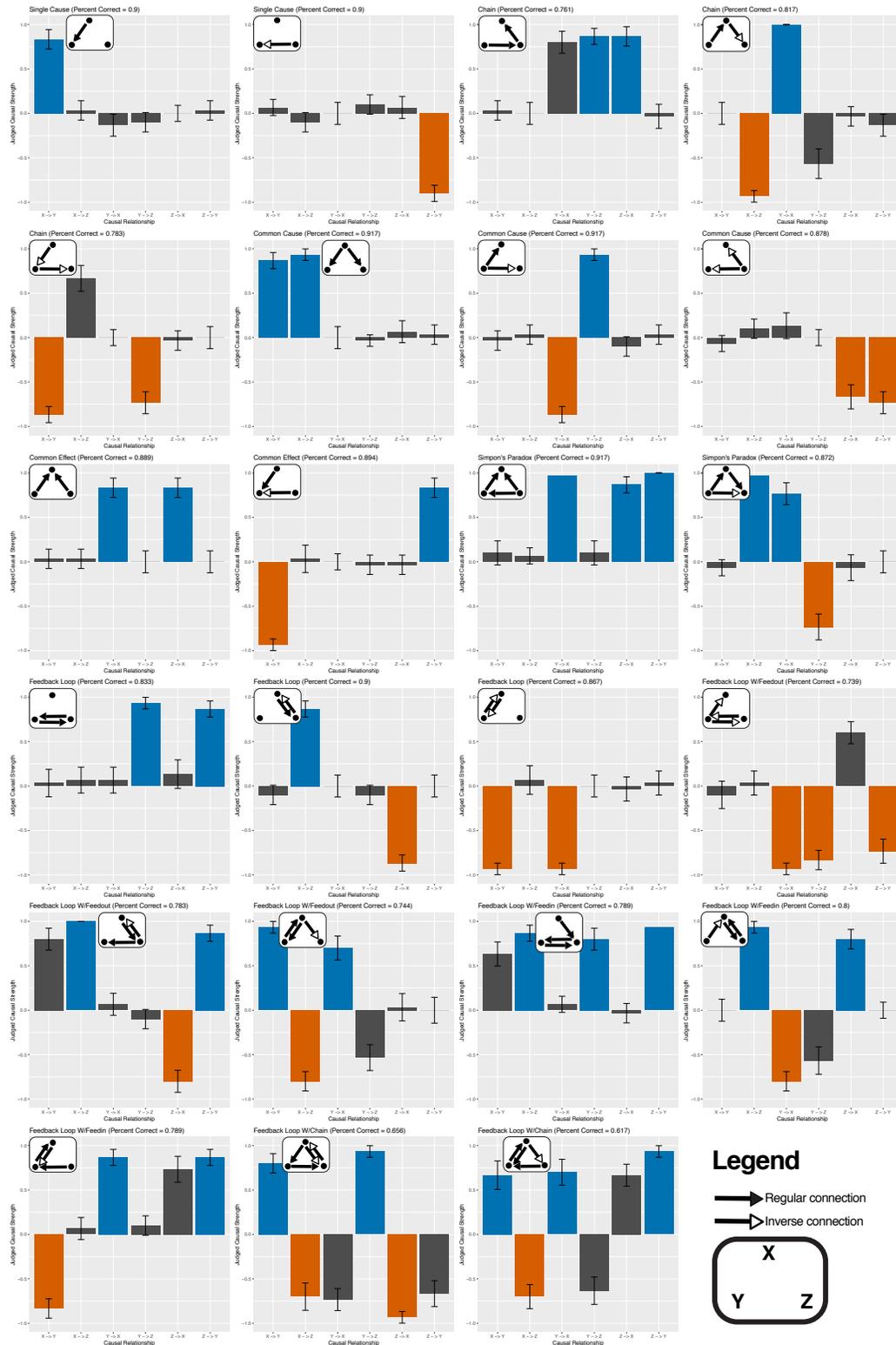


Figure 8. Participant judgments of causal relationships for all 23 tested networks. Blue, gray, and red bars correspond to regular, absent, and inverse connections, respectively. Bars represent the mean θ reported by participants, where regular = 1, none = 0, and inverse = -1. Error bars denote 95% confidence intervals.