Cognition 179 (2018) 266–297

Contents lists available at ScienceDirect

Cognition

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

# Successful structure learning from observational data

Anselm Rothe[a,*], Ben Deverett[b], Ralf Mayrhofer[c], Charles Kemp[d,1]

[a] *Department of Psychology, New York University, NY 10003, United States*
[b] *Department of Molecular Biology and Princeton Neuroscience Institute, Princeton University, NJ 08544, United States*
[c] *Department of Psychology, University of Göttingen, Germany*
[d] *Department of Psychology, Carnegie Mellon University, PA 15213, United States*

ARTICLE INFO

*Keywords:*
Causal structure learning
Causal reasoning
Bayesian modeling

ABSTRACT

Previous work suggests that humans find it difficult to learn the structure of causal systems given observational data alone. We identify two conditions that enable successful structure learning from observational data: people succeed if the underlying causal system is deterministic, and if each pattern of observations has a single root cause. In four experiments, we show that either condition alone is sufficient to enable high levels of performance, but that performance is poor if neither condition applies. A fifth experiment suggests that neither determinism nor root sparsity takes priority over the other. Our data are broadly consistent with a Bayesian model that embodies a preference for structures that make the observed data not only possible but probable.

## 1. Introduction

Causal networks have been widely used as models of the mental representations that support causal reasoning. For example, an engineer's knowledge of the local electricity system may take the form of a network in which the nodes represent power stations and the links in the network represent connections between stations. Causal networks of this kind may be learned in several ways. For example, an intervention at station A that also affects station B provides evidence for a directed link between A and B. Networks can also be learned via instruction: for example, a senior colleague might tell the engineer that A sends power to B. Here, however, we focus on whether and how causal networks can be learned from observational data. For example, the engineer might observe that A and B both have voltage spikes on some occasions, that B alone has voltage spikes on others, but that A is never the only station with voltage spikes (Fig. 1). Based on these observations alone, the engineer might infer that A sends power to B.

The problem in Fig. 1 is an instance of *structure* learning because it requires a choice between two distinct graph structures: one in which A sends a link to B and the other in which B sends a link to A. Structure learning can be distinguished from *parameter learning* problems that require inferences about the properties of links in a known causal structure (Danks, 2014; Jacobs & Kruschke, 2011). For example, an engineer who knows that station A sends a link to station B might need to learn about the fidelity with which signals at A are transmitted to B. Causal parameter learning is often studied experimentally using paradigms in which a focal effect is clearly distinguished from a set of potential causes, and the learning problem is to infer the strength of the relationship between each candidate cause and the effect (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Sloman, 2005). Here, however, we focus on structure learning problems in which the variables are not presorted into potential causes and effects.

A consensus has emerged that people find causal structure learning to be difficult or impossible given observational data alone. For example, Fernbach and Sloman (2009) cite results obtained by Steyvers, Tenenbaum, Wagenmakers, and Blum (2003), Lagnado and Sloman (2004), and White (2006) to support their claim that "observation of covariation is insufficient for most participants to recover causal structure" (p. 680). Here we challenge this consensus by identifying two conditions that enable successful structure learning from observational data alone. The first condition is causal determinism, and is satisfied if each variable is a deterministic function of its direct causes. The second condition is root sparsity, and is satisfied if each observation is the outcome of a single root cause. Both conditions simplify the structure-learning problem by reducing the number of possible explanations for a given set of observations.

Determinism and root sparsity have both previously been discussed in the literature on causal reasoning. Several lines of research suggest that people tend to assume that causes are deterministic or near-deterministic (Frosch & Johnson-Laird, 2011; Lu et al., 2008; Schulz & Sommerville, 2006; Yeung & Griffiths, 2015), and this assumption has informed previous studies of structure learning (Mayrhofer &

[*] Corresponding author at: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, United States.
*E-mail address:* anselm@nyu.edu (A. Rothe).
[1] Present address: School of Psychological Sciences, University of Melbourne, Australia.

https://doi.org/10.1016/j.cognition.2018.06.003
Received 24 March 2017; Received in revised form 31 May 2018; Accepted 4 June 2018

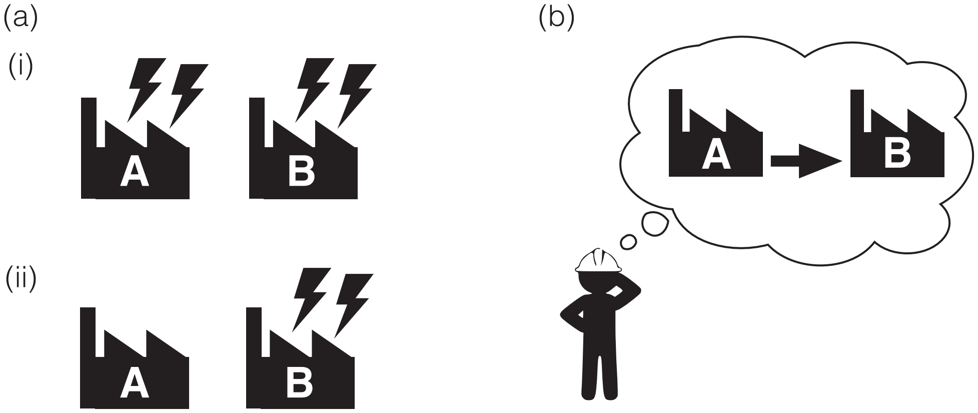


**Fig. 1.** Learning the causal structure of a power network given observations alone. (a) When voltage spikes are observed, either (i) stations A and B both have voltage spikes or (ii) B alone has voltage spikes. (b) These observations support the inference that station A sends power to station B.

Waldmann, 2011; Mayrhofer & Waldmann, 2016; White, 2006). Our work is related most closely to a previous study by White (2006), who asked participants to learn the structure of deterministic causal systems from observational data alone. White's task proved to be difficult, and performance was poor even when White gave his participants explicit instructions about how to infer causal structure from observational data. In contrast, we find that our participants are reliably able to infer the structure of deterministic causal systems.

Although "root sparsity" is our own coinage, this term is related to a cluster of existing ideas. Some work on causal attribution suggests that people tend to prefer explanations that invoke a single root cause (Chi, Roscoe, Slotta, Roy, & Chase, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017), although Zemla, Sloman, Bechlivanidis, and Lagnado (2017) report the opposite finding. Many studies of causal parameter learning consider cases in which there are two potential causes of an effect: a focal cause and a background cause. In this setting learners seem to expect that exactly one of these potential causes is strong (Lu et al., 2008). Mayrhofer and Waldmann (2015) explore a related idea in their work on prior expectations in structure learning. One of the priors that they consider captures the idea that an effect has a single cause. The notion of root sparsity is also consistent with studies of structure learning that focus on the role of interventions. Several researchers in this literature suggest that people tend to succeed only when interventions are not accompanied by spurious changes. If this condition holds then all changes observed following an intervention can be traced back to a single root cause – that is, to the intervention (Fernbach & Sloman, 2009; Lagnado & Sloman, 2004). Rottman and Keil (2012) show that the same condition supports structure learning from observational data if the temporal sequence of the observations is known.

Our primary goal is to explore the extent to which determinism and root sparsity allow people to succeed at structure learning. We find that people perform well when determinism and root sparsity both apply, and that either condition alone is sufficient to produce high levels of performance. To help us understand our participants' inferences, we compare these inferences to the predictions of several computational models. We initially focus on a model that we refer to as the Bayesian structure learner, or the BSL for short. The BSL serves as a normative benchmark that helps to evaluate the extent to which people succeed at structure learning. Previous discussions of structure learning have also considered Bayesian benchmarks, but Fernbach and Sloman (2009) suggest that there is "little reason to treat them as descriptively correct" (p. 681). In our setting, however, we find that people's inferences align closely with the predictions of our Bayesian model in many cases.

The BSL model contrasts with previous statistical accounts of structure learning that are sensitive to patterns of conditional independence between variables (Pearl, 2000; Spirtes, Glymour, & Scheines, 2001). Like several previous authors (Fernbach & Sloman, 2009; Mayrhofer & Waldmann, 2011), we believe that models that track patterns of conditional independence are often too powerful to capture inferences made by resource-bounded human learners. The BSL model uses statistical inference in a different way, and relies on a computation that assesses how much of a coincidence the available data would be with respect to different possible structures. It is therefore possible that people rely on a similar kind of statistical computation when approaching structure learning problems.

## 2. Four classes of causal networks

The causal systems that we consider are simple activation networks. Each network can be represented as a graph which may include cycles. Each node in the graph can be active or inactive, and the edges in the graph transmit activation from one node to another.

This paper will consider four qualitatively different classes of causal networks that are summarized in Table 1. The causal links in a network may be deterministic (D) or probabilistic (P), and root causes may be sparse (S) or non-sparse (N), producing a total of four possibilities that we refer to as classes DS, DN, PS, and PN.

Fig. 2a shows an example of activation spreading over a network from class DS. At stage i, node A activates spontaneously. At stage ii, node A has activated nodes B and C. At stage iii, node B has activated node D, and the network has reached a stable end state. The links in the network are deterministic, which means that they always succeed in transferring activation from one node to another. Root causes are sparse, which means that at most one node activates spontaneously per trial. As a result, each end state is the consequence of a single root cause. For example, the end state in Fig. 2a.iii is the consequence of the initial activation of node A.

Fig. 2b shows a network for which root causes are non-sparse. At

**Table 1**
Four classes of causal networks. For each class, the number of possible causal histories for a network with $n$ nodes and $l$ links is shown.

| Causal strength | Number of root causes | |
|---|---|---|
| | 1 | ≥ 1 |
| Deterministic | Class DS (deterministic and sparse)<br>Experiment 1<br>$n$ | Class DN (deterministic and non-sparse)<br>Experiment 2<br>$2^n$ |
| Probabilistic | Class PS (probabilistic and sparse)<br>Experiment 3<br>$n(2^l-1)$ | Class PN (probabilistic and non-sparse)<br>Experiment 4<br>$2^n(2^l-1)$ |

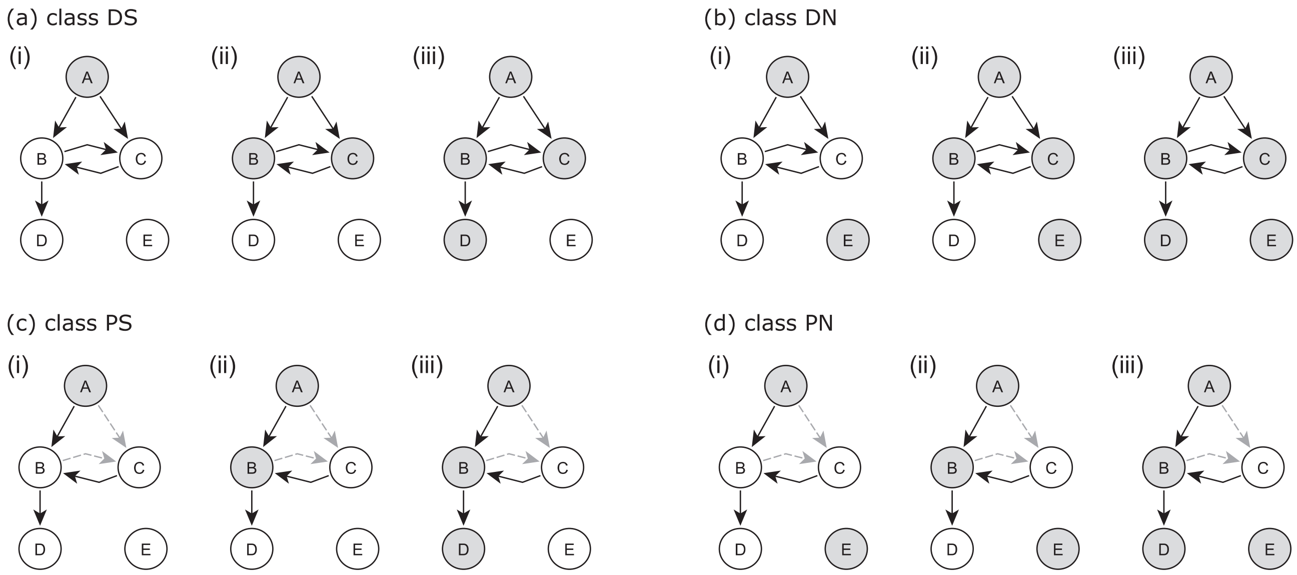


**Fig. 2.** Activation spreading over example networks from each class. In each case, panel (i) shows a start state and panel (iii) shows a stable end state. In our structure-learning task, the arrows were hidden and participants were asked to infer the structure of a network given a number of stable end states generated over that network.

stage i, there are two spontaneous activations. Activation subsequently spreads through the network, producing a stable end state in which all nodes are active.

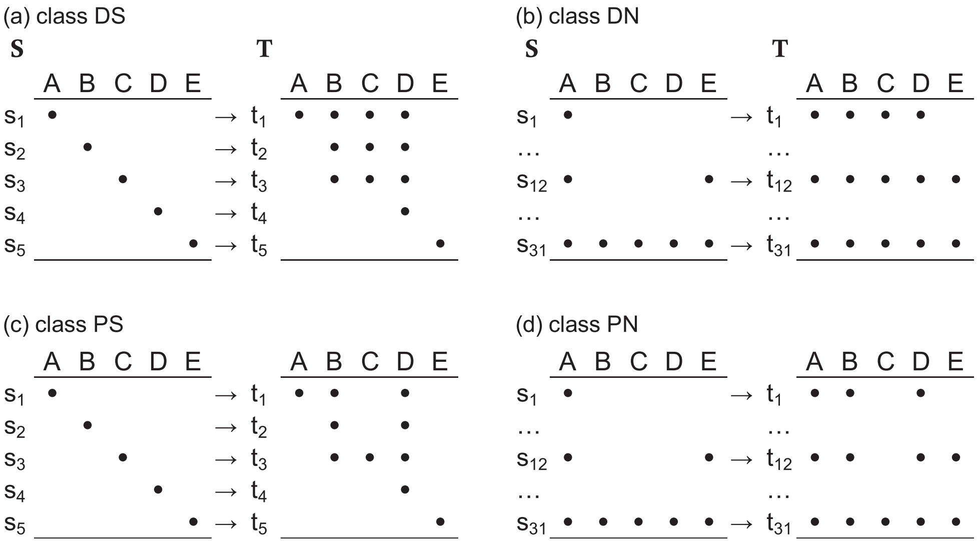
Fig. 2c shows a network with probabilistic links. On any given trial, some links may be active (shown as black arrows) but others may be inactive (shown as dashed gray arrows). Given a start state along with information about which links are active, we assume that causal activation propagates deterministically along the active links. For example, in Fig. 2c, node A activates node B but A fails to activate C because the link joining these nodes is inactive. The probabilistic nature of the links can be captured using exogenous factors that determine which links are active on a given trial. As discussed later, these exogenous factors can be represented as variables—one for each link—and the causal system can be represented as a functional causal model (Pearl, 2000) in which exogenous variables are stochastic but all other variables are deterministic functions of their parents.

Fig. 2d completes the set, and shows a network for which causal links are probabilistic and root causes are non-sparse. Nodes A and E simultaneously activate in stage i, but activation fails to reach node C because both links directed towards this node are inactive.

Fig. 2 shows only one end state for each of the networks shown. The matrix notation in Fig. 3 is a convenient way to represent the full set of end states for a given network. The rows of matrix $S$ represent the possible start states and each row of matrix $T$ represents the corresponding end state. For example, in Fig. 3b, $s_{12}$ and $t_{12}$ represent the start and end states shown in Fig. 2b.

Although Fig. 2 shows how activation propagates over the four networks, we consider the learning problem in which the end state alone is observed and the learner must infer the structure of the network. We will be especially interested in comparing patterns of performance across the four classes of networks, and learning whether people are able to reliably solve the structure learning problem for any of the four classes.

Although our four classes of networks represent a range of possibilities, the activation networks that we consider are constrained in important ways. For example, they include only generative causal relationships, and they assume that multiple causes combine according to an OR function. Given the state of the literature, our most pressing question is whether there are any classes of networks for which people reliably succeed at structure learning. Instead of aiming for



(a) class DS

| S | A | B | C | D | E | | T | A | B | C | D | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1$ | ● | | | | | → | $t_1$ | ● | ● | ● | ● | |
| $s_2$ | | ● | | | | → | $t_2$ | | ● | ● | ● | |
| $s_3$ | | | ● | | | → | $t_3$ | | ● | ● | ● | |
| $s_4$ | | | | ● | | → | $t_4$ | | | | ● | |
| $s_5$ | | | | | ● | → | $t_5$ | | | | | ● |

(b) class DN

| S | A | B | C | D | E | | T | A | B | C | D | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1$ | ● | | | | | → | $t_1$ | ● | ● | ● | ● | |
| … | | | | | | | … | | | | | |
| $s_{12}$ | ● | | | | ● | → | $t_{12}$ | ● | ● | ● | ● | ● |
| … | | | | | | | … | | | | | |
| $s_{31}$ | ● | ● | ● | ● | ● | → | $t_{31}$ | ● | ● | ● | ● | ● |

(c) class PS

| | A | B | C | D | E | | | A | B | C | D | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1$ | ● | | | | | → | $t_1$ | ● | ● | | ● | |
| $s_2$ | | ● | | | | → | $t_2$ | | ● | | ● | |
| $s_3$ | | | ● | | | → | $t_3$ | | ● | ● | ● | |
| $s_4$ | | | | ● | | → | $t_4$ | | | | ● | |
| $s_5$ | | | | | ● | → | $t_5$ | | | | | ● |

(d) class PN

| | A | B | C | D | E | | | A | B | C | D | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1$ | ● | | | | | → | $t_1$ | ● | ● | | ● | |
| … | | | | | | | … | | | | | |
| $s_{12}$ | ● | | | | ● | → | $t_{12}$ | ● | ● | | ● | ● |
| … | | | | | | | … | | | | | |
| $s_{31}$ | ● | ● | ● | ● | ● | → | $t_{31}$ | ● | ● | ● | ● | ● |

**Fig. 3.** Activation matrices of start states (S) and end states (T) for the four classes. For classes DN and PN only three of the 31 possible start states are shown. For classes PS and PN the end states depend on which links in the network are inactive: the end states shown here are for the case in Fig. 2 for which two links are inactive.

comprehensive coverage of the space of causal networks, we aimed to work with causal systems that would give people the greatest chance of performing well at structure learning.

We felt that working with DS networks (i.e. deterministic networks with sparse root causes) would make the structure learning problem as easy as possible. The remaining three network classes were created by relaxing one or both of the characteristic assumptions of DS networks. The determinism and root-sparsity assumptions both simplify the structure learning problem because they reduce the number of end states that can be generated over a given network. For example, if node A sends a link to B, the two assumptions rule out states in which A is active but B is not.

Our idea that DS networks might make the structure learning problem especially tractable is consistent with the literature on causal learning. Recent studies suggest that children and adults both tend to assume that causal relationships are deterministic (Lu et al., 2008; Schulz & Sommerville, 2006; Yeung & Griffiths, 2015), and it is natural to expect that people are most likely to succeed at structure learning when reasoning about systems that match this assumption. At least one previous account of structure learning directly incorporates the determinism assumption. Mayrhofer and Waldmann (2011) propose that people solve structure learning problems by identifying the structure that minimizes the number of cases in which a cause is present but an effect is absent. As these authors note, this “broken link” heuristic corresponds to the expectation that causes are near-deterministic.

Some previous work also suggests that children and adults both tend to assume that root causes are sparse (Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017). For example, Lombrozo (2007) found that people preferred to attribute an observation to a single root cause, even in cases in which a two-cause explanation was more probable. Root sparsity is also consistent with the work of Chi et al. (2012), who suggest that people’s explanations of both scientific and everyday processes tend to invoke a single initiating or triggering event. This emphasis on root sparsity, however, is challenged by Zemla et al. (2017), who report that the perceived quality of an explanation is correlated with the number of root causes that it invokes. The factors that moderate the explanatory importance of root sparsity are not yet clear, but Johnson, Valenti, and Keil (2017) suggest that root sparsity carries more weight when reasoning about deterministic causal systems than when reasoning about stochastic systems.

The expectation that root causes are sparse and that causal links are deterministic is reminiscent of the work of Lu et al. (2008) on “sparse and strong” priors for causal learning. Importantly, however, our notion of sparsity is different from theirs. Their notion is formulated in terms of *type* causation, and captures the expectation that each node in a causal graph is expected to have at most one strong cause.[2] Our notion connects more closely with *token* causation, and captures the idea that each observed pattern of activation is expected to have a single root cause. The two notions are not equivalent, and neither can be reduced to the other. For example, the activation network in Fig. 2a is inconsistent with their notion of sparsity, because A and B are both strong causes of C. This network, however, is consistent with our notion of root sparsity as long as spontaneous activations are very rare, which means that each observed end state will result from the activation of a single node.

Although our notion of root sparsity connects with token causation, the problem of structure learning is typically addressed with respect to type causation and we follow in this tradition. More precisely, we consider how a learner could infer a causal structure over a set of variables after observing multiple patterns of activation over these variables. These patterns are assumed to be generated over a single, underlying network, and the edges in this network capture type causation. Root sparsity applies if each observed pattern of activation is the consequence of activating a single node in the network.

Much of the literature on causal learning considers directed acyclic networks, but our four classes of networks include networks with cycles. People often generate cycles when asked to draw causal networks (Kim, Luhmann, Pierce, & Ryan, 2009), and cycles seem especially natural in the context of our activation networks. If desired, these activation networks could be represented as dynamic Bayesian networks where the cycles are unrolled in time (Rehder & Martin, 2011). As discussed later, our activation networks can also be represented using functional causal models, which allow the possibility of cycles (Pearl, 2000).

From one perspective, allowing for cycles increases the difficulty of structure learning by expanding the size of the hypothesis space of possible structures. Given that most previous studies of structure learning rule out cycles and find structure learning to be difficult, it would be surprising to find that people succeed at structure learning when cycles are allowed. From another perspective, ruling out cycles would make our experiments more difficult by compromising the naturalness of the structure-learning task. In our setting, there is no particular reason why cycles cannot occur, and asking participants to generate acyclic structures would therefore amount to asking them to operate under an arbitrary constraint.

## 3. Bayesian structure learning

We now describe a Bayesian approach to the problem of structure learning. The primary purpose of our Bayesian framework is to provide a benchmark for assessing how well people learn structures from the four classes just described.

Suppose that we observe a data set $D$ generated from an unknown network $G$. Our framework can be applied to problems based on all four of the network classes in Fig. 2, but we will initially assume that the unknown network belongs to class DS: in other words, that causal relationships are deterministic and that root causes are sparse.

Data set $D$ can be formulated as a matrix, where each row $d_i$ represents an end state sampled from network $G$. After observing $D$, the posterior distribution over networks $G$ is

$$P(G|D) \propto P(D|G)P(G) = \left[\prod_i P(d_i|G)\right] P(G) \tag{1}$$

where we have assumed that the observations $d_i$ are independently generated over network $G$. To complete the model we need to specify the likelihood term $P(d_i|G)$ and the prior $P(G)$. For now, we use a uniform prior $P(G)$, and compare two possible versions of the likelihood.

The BSL assumes that each observation $d_i$ is an end state that resulted from a start state randomly sampled from a prior $P(s)$ on start states. The resulting likelihood can be computed by summing over possible start states $s_j$:

$$P(d_i|G) = \sum_j P(d_i|G, s_j)P(s_j) \tag{2}$$

Because root causes are sparse, each start state includes a single active node, and we assume that the prior on start states $P(s_j)$ is uniform. $P(d_i|G, s_j)$ is either 1 or 0 depending on whether $d_i$ is the end state associated with start state $s_j$: that is, $P(d_i|G, s_j) = 1$ if and only if initializing the network in state $s_j$ and allowing activation to propagate through the network results in the stable state $d_i$. We refer to Eq. (2) as the *graded* likelihood because it captures the idea that some observations are more typical than others. For example, given the network in Fig. 2a, the stable state ABCD can be produced in one way (A must be the root cause) but the stable state BCD can be produced in two ways (either B or C can be the root cause). For this network, the graded likelihood captures the idea that BCD is more probable than any other

[2] Lu et al. (2008) focus on graphs with two nodes, and do not explicitly say how their notion of sparsity extends to arbitrary graphs. The idea that each node has at most one strong cause seems like the most natural generalization to us.

observation.

Although the BSL relies on the graded likelihood, other Bayesian models can be formulated by making different assumptions about the likelihood. One possible alternative is a binary likelihood that considers only whether observation $d_i$ is consistent with $G$:

$$P(d_i|G) = \begin{cases} 1 & \text{if } d_i \text{ is consistent with } G \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Observation $d_i$ is consistent with $G$ if $d_i$ is an end state that results from at least one valid start state. For the network in Fig. 2a, the binary likelihood assigns the same probability to observations BCD and ABCD.

Combining the binary likelihood with a uniform prior $P(G)$ produces a model that we refer to as the *broken link* model. The broken link model computes a posterior distribution $P(G|D)$ that assigns equal probability to all graphs $G$ that are consistent with the data, and zero probability to all remaining graphs. The model is consistent with the broken link heuristic described by Mayrhofer and Waldmann (2011), which assesses how well graph $G$ accounts for data $D$ by counting the number of times that a parent node is active and a child node is inactive. For networks belonging to class DS or DN, a graph is deemed consistent with data $D$ if and only if the graph has a broken link count of zero. When applied to these classes of networks, the broken link model is therefore equivalent to a model which assumes that people choose a graph that minimizes the broken link count, and that people are indifferent among graphs that satisfy this criterion.

From a normative perspective, the two versions of the likelihood do not have equal status. Our experiments explore settings in which each observation $d_i$ is generated from a causal process that unfolds from a start state, and observations that can be generated from multiple different start states should be encountered more frequently. The BSL model is consistent with this notion and the broken link model is a foil that will be useful for exploring the extent to which people's inferences are consistent with the graded likelihood.

To see how the models differ, consider a three-node problem where $D$ includes 6 observations and where each observation indicates that nodes A, B and C are all active. The data are consistent with multiple structures, but here we focus on two: a fully connected graph, and a causal chain where A sends an arrow to B and B sends an arrow to C (graphs 64 and 10 in Fig. D.1 in Appendix D). Both structures are consistent with the data, and as a result the broken link model considers them to be equally probable. In contrast, the BSL recognizes that the data are not typical of the chain, and therefore assigns higher probability to the fully connected structure. Explaining the data in terms of the chain requires the assumption that A spontaneously activated 6 times in succession, which seems like a big coincidence. The BSL recognizes that the connected structure provides a better explanation, because in this case all nodes end up active regardless of which node activates first.

### 3.1. Allowing for non-sparse root causes and probabilistic links

Thus far we have described how the BSL and broken link models can be applied to the problem of learning an unknown network drawn from class DS, where causal links are deterministic and there is only one root cause. Both models, however, can also be applied when causal links are probabilistic and root causes are non-sparse, as occurs for classes DN, PS, and PN. We now discuss how the likelihood terms are adjusted for these classes.

Suppose that the unknown network $G$ is known to belong to class DN. This class differs from class DS in that two or more root causes may simultaneously occur (cf. Table 1). The graded and binary likelihoods in Eqs. (2) and (3) are now adjusted to allow for start states with more than one active node. The prior $P(s_j)$ on start states used by the graded likelihood is formulated in terms of a background-rate parameter $b$ that captures the probability that any given node will be spontaneously active in the start state. The prior over start states is therefore

$$P(s_j) \propto \begin{cases} b^a(1-b)^{n-a} & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{cases} \tag{4}$$

where $a$ denotes the number of active nodes in $s_j$ and $n$ denotes the total number of nodes. Consistent with our assumption that a start state must include at least one active node, the prior assigns zero probability to the state in which all nodes are inactive.

Suppose now that the unknown network $G$ is known to belong to class PS. This class differs from class DS in that causal links might be inactive and fail to transmit activation. The graded and binary likelihoods in Eqs. (2) and (3) must be adjusted again to allow for the possibility of inactive links. For each graph $G$, we consider all possible variants $G_v$ produced by inactivating zero or more of the links in $G$. For example, Fig. 2c shows a variant of the underlying five node-network in which the links A → C and B → C are inactive. The graded likelihood now incorporates a sum over all possible variants $G_v$ of graph $G$:

$$P(d_i|G) = \sum_{G_v}\left[\sum_j P(d_i|G_v, s_j)P(s_j)\right]P(G_v|G) \tag{5}$$

The prior $P(G_v|G)$ on graph variants is formulated using a failure rate $f$, which captures the probability that any given link will be inactive. The prior $P(G_v|G)$ is therefore

$$P(G_v|G) = (1-f)^{l-i}f^i \tag{6}$$

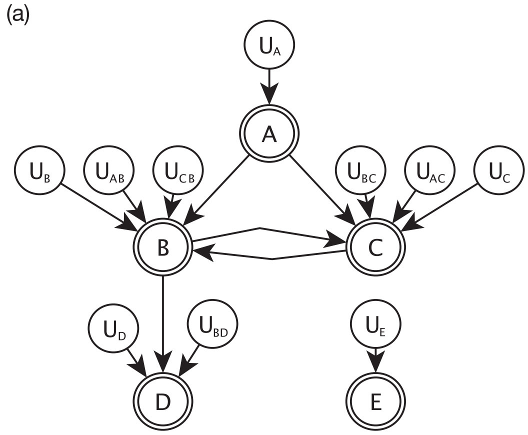
where $l$ denotes the number of links in $G$ and $i$ the number of inactive links in $G_v$.

An alternative way to derive the likelihood $P(d_i|G)$ is to work with functional causal models. Fig. 4 shows a functional causal model that corresponds to the activation network shown in Fig. 2. Exogenous variables have been introduced to capture the factors that determine whether a link is active on a given trial. For example, $U_{BD}$ is a binary variable that determines whether or not the link from B to D is active. The functional model also includes exogenous binary variables such as $U_D$ that determine whether a node (in this case, $D$) is active in the start state. For class PS, the base rates of the five variables $U_A$ through $U_E$ should be set to a very small value to capture the idea that the start state almost certainly includes at most one active node. For class PN, the base rates can be set to a value such as 0.5. After introducing the full set of exogenous variables, each node in the activation network can be represented as a deterministic function of its parents. For example, Fig. 4b shows that node D is active if $U_D$ is true (meaning that D is active in the start state) or if $U_B$ and $U_{BD}$ are both true (meaning that B is active and the link from B to D is active). Given a functional model such as Fig. 4, the likelihood $P(d_i|G)$ is computed in the standard way by summing out over all possible settings of the exogenous variables. Summing out over the variables $U_A$ through $U_E$ is equivalent to summing out over start states $s_j$ in Eq. (5), and summing out over the variables $U_{AB}$ through $U_{BD}$ is equivalent to summing out over graph variants $G_v$.

### 3.2. Relative difficulty of the four classes of networks

We suggested earlier that structure learning might be especially easy when reasoning about deterministic systems with sparse root causes (i.e. systems from class DS). One way to arrive at this conclusion is to consider the number of causal histories for a network: that is, the number of distinct ways in which the network can generate observations. Our experiments focus on networks with three nodes, each of which may have up to 6 edges. There are 64 networks and the full space is shown in Fig. D.1. This hypothesis space of networks is the same for the four different classes, but the classes differ with respect to the number of causal histories for a given network.

The *causal history set* for a network (i.e. the set of all causal histories) can be constructed by pairing every possible start state $s$ with every possible graph variant $G_v$. For example, consider a three-node network



(b)

| $U_D$ | B | $U_{BD}$ | $D = f_D(U_D, B, U_{BD})$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |

**Fig. 4.** A functional causal model that represents the PS network in Fig. 2c. (a) The network includes exogenous variables such as $U_D$, which determines whether node D is active in the start state, and $U_{BD}$, which determines whether the link from B to D is active. Variables A through E have double boundaries, indicating that they are deterministic functions of their parents. (b) The conditional probability distribution for node D in the functional causal model. Node D is active if variable $U_D$ is active or if B and $U_{BD}$ are both active.

from class DS. The network has three possible start states, and a single graph variant (the variant that includes no inactive links). As a result, the causal history set for the network includes three possible histories.

More generally, consider a network with $n$ nodes and $l$ links. If the network has sparse root causes, then the number of start states is $n$. If root causes are not sparse, then the number of start states is $2^n-1$. If the network's links are deterministic, then only one graph variant $G_v$ is possible. If the network's links are probabilistic, then $2^l$ graph variants are possible. Combining all possible start states with all possible graph variants produces the causal history set, and Table 1 shows the size of this set for each of the four network classes.

In general, networks with large causal history sets may be difficult to reason about because there are many different ways in which observations can be generated over these networks. This expectation supports predictions about the relative difficulty of structure learning for the four network classes. For simplicity, assume that $n$ and $l$ are roughly the same. This assumption holds for the three node networks in Fig. D.1, which have three links on average. Given this assumption, the causal history counts in Table 1 predict that the difficulty order for the four classes is

$$\mathrm{DS} < \mathrm{DN} \ < \ \mathrm{PS} < \mathrm{PN} \tag{7}$$

There are other ways to derive a difficulty order over the four classes. For example, one could consider a Bayesian model such as the BSL and assess how often the model correctly reconstructs the true structure when given problems based on the four classes. Under some assumptions, it seems possible that the relative difficulties of classes DN and PS will be reversed. We expect, however, that all sensible derivations will agree that structure learning is easiest for class DS and most difficult for class PN.

## 4. Overview of experiments

We designed a series of experiments in which participants learned networks from the four classes in Table 1. Our primary goal was to explore whether people would perform well at structure-learning when reasoning about systems with deterministic causal links and systems that satisfy the root sparsity condition. Previous research suggests that people often perform poorly on structure-learning tasks given observational data alone (Steyvers et al., 2003; White, 2006), but we suspected that participants might perform relatively well if they could assume causal determinism and root-sparsity.

In all of our experiments, participants observed end states generated over an unknown network, then drew a graph to indicate their inferences about the structure of the network. Experiments 1 through 4 used networks belonging to classes DS, DN, PS and PN respectively. The sequence of experiments therefore follows the difficulty order predicted in Eq. (7). The instructions for each experiment informed participants about the class of networks that they should consider. For example, participants in Experiment 1 (class DS) were given reason to believe that the unknown network had deterministic links, and that each observed end state had a single root cause.

The results of Experiments 1 through 4 suggest that causal determinism and root sparsity both enable successful structure learning from observational data. Experiment 5 explored whether one of these factors is more fundamental than the other. Participants were given cases that could only be explained by abandoning either the determinism assumption or the root sparsity assumption, and we asked whether participants tended to agree about which assumption to abandon and which to preserve.

A secondary goal of our experiments was to evaluate the BSL and to compare it to the broken link model. These two models capture different theories about the assumptions that people bring to structure learning, and comparing these models provides a way to characterize the assumptions that people actually make. The instructions for each experiment included examples of activations that allowed people to estimate any relevant numerical parameters including the background rate (Eq. (4)) and the failure rate (Eq. (6)). When implementing the models, numerical parameters were set to maximum likelihood values based on the introductory examples. No free parameters were incorporated at any stage, and the models can therefore be compared on equal terms.

## 5. Experiment 1: Deterministic links, one root cause

Because structure learning given observational data alone is traditionally thought to be difficult, Experiment 1 explores the network class (DS) that makes this problem as easy as possible. Our task is based on activation networks with three nodes, and we included structure-learning problems based on all such networks that are qualitatively different. The space of these networks includes common-cause structures and common-effect structures, both of which have proved difficult to learn in previous experiments (Steyvers et al., 2003; White, 2006).

Among these previous experiments, our experimental paradigm is closest to the work of White (2006), who ran a structure-learning task based on deterministic causal networks. White found that performance was relatively poor, but this result may be due in part to the nature of his experimental materials. In particular, White's primary cover story involved changes in the populations of animal species over time, and this emphasis on temporal order may have made it difficult for people to solve the problem of structure learning. Our experiment asked people to reason about networks of particle detectors, and we did our best to introduce these networks and the structure-learning problem in a way that was as intuitive as possible.

## 5.1. Method

### 5.1.1. Participants

36 members of the Carnegie Mellon University (CMU) community participated in exchange for pay or course credit. In this and all following experiments, informed consent was obtained from each participant.

### 5.1.2. Materials

The causal systems in this experiment were described as networks of particle detectors. Particle detectors (i.e. nodes) were shown as rectangles. The interior color of a detector indicated whether it was inactive (gray) or active (green).[3] Links between detectors (i.e. causal relationships) were shown as black arrows.

In each block of the experiment, participants observed a set of end states generated over an unknown network, and were asked to infer the structure of the network. There were 32 blocks in total, and each involved a network with three nodes. The blocks were systematically constructed to cover the space of observations that can be generated from three node networks in class DS. Here we describe the method for generating blocks in a way that will generalize to subsequent experiments.

We previously defined the *causal history set* as a set that includes all ways in which observations can be generated over a given network. The causal history set can be used to generate the *characteristic observation set* for a network (or characteristic set for short), which includes one observation corresponding to each possible causal history. Consider, for example, the fully connected DS network (number 64 in Fig. D.1). There are three possible causal histories for the network (either A, B or C can be active in the start state), and each causal history leads to an end state in which all nodes are active, which means that the characteristic set for the network is {ABC, ABC, ABC}. As a second example, the characteristic set for the empty DS network is {A, B, C}.

Each characteristic set induces a *characteristic distribution* over observations. For example, the characteristic distribution for the fully connected DS network assigns probability 1 to the observation ABC. The characteristic distribution for the empty DS network assigns probability 1/3 to each element in {A, B, C}.

Different networks often have the same characteristic set, and therefore the same characteristic distribution. For example, a DS network that corresponds to a three-node cycle has the same characteristic distribution as the fully-connected DS network mentioned above. Different networks may also have characteristic distributions that are not the same, but are identical up to relabeling of the nodes. For example, the network with a single link from A → B and the network with a single link from A → C generate characteristic distributions that are not the same, but are identical up to relabeling. If we group characteristic distributions into classes that are equivalent up to relabeling, the 64 three node DS networks generate 9 qualitatively different characteristic distributions.

These nine characteristic distributions were the basis for the blocks in Experiment 1. 18 of these blocks represented the characteristic distributions using three observations each. These 18 blocks included two instances of each of the 3-observation blocks shown in Table 2. An additional 9 blocks represented the characteristic distributions using six observations each, and the remaining 5 blocks each included one or two observations each and are listed in Table B.1.

A network is listed as a "generating structure" in Table 2 if its characteristic distribution matches a given block. There are multiple generating structures for some blocks. For example, there are several qualitatively different structures that can only generate the observation ABC (Block 9). When analyzing our data, we will treat the generating structures in Table 2 as the "correct" responses for each block. It is straightforward to show that the generating structures for a block are precisely those that emerge as the most probable structures for that block according to the BSL model.

### 5.1.3. Procedure

Participants interacted with a custom-built graphical interface that presented them with end-states generated over a network with unknown structure, and allowed them to record their inferences by drawing causal links between the detectors in a network (see Fig. 5). After an introduction to the experiment, the 32 blocks were presented in random order, and each block included an observation phase and a test phase.

*5.1.3.1. Introduction.* Participants were shown the five-node network in Fig. 2 and told that the boxes were detectors that detect a rare type of particle called the mu particle. The arrows were described as connections between these detectors. Participants were told that an active detector always activates all detectors that it points to. To reinforce this information, participants were given an example like Fig. 2a that showed a single detector activating and activation propagating over the network in a series of steps. Once the network had reached a stable end state, this state was added to an "observation panel" on the left of the screen. The system was then reset—that is, all detectors were set to inactive, but the links between detectors remained.

Participants then observed two similar examples of activation spreading over the network. The end states for both examples were added to the observation panel, making a total of three observed states.

Next, participants were informed that during the experiment, the arrows in the network would be hidden and that they would need to figure out which arrows existed. The arrows were subsequently removed from the network shown on screen, and the same three examples were presented again. Importantly, only the end states were shown this time, consistent with the blocks to follow.

*5.1.3.2. Observation phase.* The observation phase for each block began with three detectors displayed on the right of the screen. The positions of the detectors were randomized but the distance between detectors was small, meaning that they formed a messy-looking pile. This design choice was intended to reinforce the fact that the initial positions of the detectors provided no information about the causal structure of the network. At any time during the block, participants could drag the detectors around with the mouse and arrange them as they liked on screen.

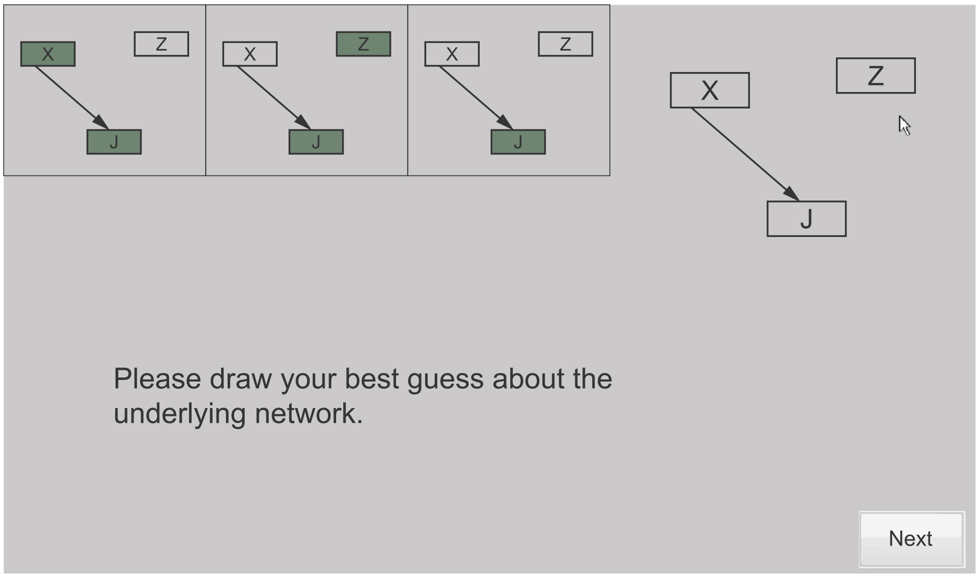
Next, the first observation for the block was shown. To minimize demands on memory, this observation was stored in the observation panel, then all detectors were set to inactive and participants were told that the network had been reset. The same procedure was used to present all remaining observations for the block. The observations within each block were presented in random order.

*5.1.3.3. Test phase.* After observing all observations in a block, participants were asked to infer the structure of the network that generated these observations. They provided their response by using the mouse to draw arrows between boxes. If desired, they could also use the mouse to delete arrows that they had previously drawn. Every drawn arrow was simultaneously added to all observations in the observation panel to make it easy for participants to see whether the

[3] In Experiment 1 we actually used red to indicate active detectors, but later experiments used green for consistency with the color scheme used to represent probabilistic links. The color used to indicate node activations did not seem to influence our results, and for simplicity the main text assumes that this color is always green. The actual color used in each experiment is shown in Table B.2.

**Table 2**
Nine blocks of three observations used in Experiment 1. The "Generating structure" column shows networks with characteristic distributions that correspond to each block. One representative is included for each class of networks that are the same up to relabeling.

| | Frequency | | | | | | | Generating structure |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC | |
| Block 1 | 1 | 1 | 1 | - | - | - | - | |
| Block 2 | - | 1 | 1 | 1 | - | - | - | |
| Block 3 | - | - | 1 | 2 | - | - | - | |
| Block 4 | 1 | 1 | - | - | - | - | 1 | |
| Block 5 | - | - | 1 | - | 1 | 1 | - | |
| Block 6 | - | - | 1 | - | - | 1 | 1 | <br> |
| Block 7 | - | - | - | 2 | - | - | 1 | <br> |
| Block 8 | - | - | 1 | - | - | - | 2 | <br> |
| Block 9 | - | - | - | - | - | - | 3 | <br> <br> <br> <br> |



Please draw your best guess about the underlying network.

**Fig. 5.** Experimental interface showing a presentation of Block 5 from Table 2. The observation panels at the top left show three observations provided during the observation phase. The learner is now asked to infer the structure of the underlying network, and has drawn a link from X to J that is duplicated in all three observation panels. The generating structure for this block includes a link from Z to J in addition to the link drawn by the learner.

structure they had drawn was consistent with all observations.

After providing their response, participants were asked to rate their confidence in their answer on a scale from 1 to 7. Participants then went on to the observation phase of the next block.

### *5.2. Results and discussion*

We focus here on responses to the 3-observation blocks shown in Table 2. Responses to the remaining blocks are summarized in Fig. D.2, and are consistent with the conclusions that we derive from the nine blocks in Table 2. Each block in Table 2 was presented twice, and all analyses here incorporate data from both block presentations. For simplicity, all analyses focus on the structures inferred by participants, and we will not discuss their confidence ratings. Data from all experiments are available online.[4]

Fig. 6 summarizes the overall results. Each scatter plot compares human responses with the predictions of one of the four models. For each of the nine blocks, we computed a human distribution over the 64 structures based on the frequency with which participants selected each structure. For example, every participant chose the empty structure (i.e. the structure without any links) for block 1, which means that the human distribution for this block assigns probability 1 to this structure and probability 0 to all remaining structures. The human distribution for a given block can be compared with a model's posterior distribution over the 64 structures. Fig. 6 combines results for all nine blocks, and each block contributes 64 data points to each scatter plot, one for each structure. For example, block 1 contributes one point at $(1,1)$ and 63 points at $(0,0)$ to the first panel of Fig. 6, because the human distribution and the BSL model both assign probability 1 to the empty structure.

Comparing the model posterior with data requires some linking hypothesis about how individuals generate their responses. Comparing the model posterior with the distribution across participants is consistent with the hypothesis that participants respond by probability matching, or sampling from the posterior. Another possible strategy is maximizing, or choosing the structure with maximum posterior probability. Although maximizing is generally taken to be the normative

[4] https://osf.io/rx8fa/.

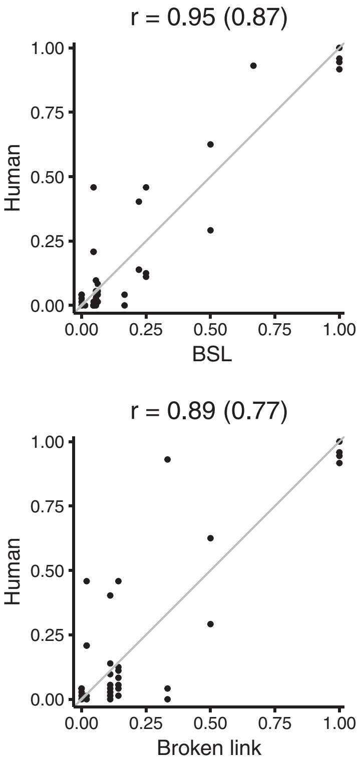


**Fig. 6.** Comparison of the complete set of human responses with model predictions for Experiment 1. In each panel the first correlation is based on the complete set of responses, and the correlation in parentheses shows the average correlation across the individual blocks of the experiment.

strategy (Eberhardt & Danks, 2011), probability matching is widely used in the literature, and provides a simple way to allow for variability in responses across participants.

The most important conclusion that can be drawn from Fig. 6 is that our participants succeeded at structure-learning given observational data alone. As suggested earlier, the BSL assigns high probability only to structures that could potentially be the true generating structure for a given block.

A more fine-grained view of the data is provided by Fig. 7, which includes the top responses for each block according to participants and the BSL model. For example, for block 1, every participant selected the empty structure, and the model assigned a posterior probability of 1 to this structure. Correct responses (i.e. responses that match the generating structures in Table 2) are enclosed in solid blue frames. In every block, the most common human response is correct. In particular, all participants discovered the common effect structure in block 3 and the common cause structure in block 6. Steyvers et al. (2003) found that these structures are difficult for learners to distinguish in a probabilistic setting, but our data suggest that they are easy to learn in our deterministic setting. Some blocks have multiple generating structures, and in these cases the top two or three responses according to participants are all correct. Overall, Fig. 7 suggests that participants performed well for each of the 9 blocks in Table 2.

The BSL and the broken link models are identical except for their likelihood function, and Fig. 6 shows that the BSL model performs better overall. In particular, the BSL model's correlation with human responses is .06 larger than that of the broken link model. To test whether this difference is statistically significant we conducted a bootstrap analysis where we sampled with replacement from the pool of participants and re-calculated the correlation between the human responses in the bootstrap sample and the model predictions. Based on 10,000 samples, we estimated a 95% confidence interval (CI) around the difference, using Efron's bias-corrected and accelerated (BCa) approach. We used the identical bootstrap routine for all confidence intervals throughout this paper.[5] For Experiment 1, the CI was $[0.05, 0.06]$, which excludes zero and thus suggests that the BSL model performed significantly better than the broken link model.

The main shortcoming of the binary likelihood is that it leads to predictions that are too diffuse. The structure preferred by participants is typically one of the most probable structures according to the broken link model, but the model often assigns the same probability to many other structures. For example, after observing ABC three times in succession, the broken link model assigns the same probability to all 51 structures that can generate the observation ABC, including causal chains over these variables. In contrast, the BSL model assigns highest probability to the 18 structures that can *only* generate the observation ABC. These 18 structures correspond to all possible relabeling of the generating structures shown in Table 2 for block 9.

Although the BSL model performs better than the broken model, its predictions are still more diffuse than the human responses. As just mentioned, the BSL model predicts that 18 different structures are equally likely after observing ABC three times, but participants overwhelmingly prefer the structures shown in Fig. 7. We return to this issue later and show how can be addressed in part by augmenting the BSL model with a prior distribution that captures a preference for symmetric structures.

Taken overall the results of Experiment 1 support two general conclusions. First, humans succeed at structure learning when causal links are deterministic and when each observation has a single root cause. To our knowledge, our data represent the first clear case of successful causal structure learning from non-temporal observational data. Our findings contrast with those of White (2006), who found that deterministic causal systems are difficult for people to learn. We return to this difference in the general discussion, and consider several factors that might help to explain it.

A second general conclusion is that structure learning in our setting cannot be adequately characterized as a search for a structure that is consistent with the observed data. Instead, people seem to be sensitive to whether a candidate structure makes the observable data not only possible but probable. The BSL model illustrates how this tendency can be captured by the likelihood of a Bayesian model, and suggests the value of the Bayesian approach to structure learning.

Given that our participants succeeded at learning the structures of networks from class DS, it is natural to ask whether determinism and root sparsity are both essential in order to enable high levels of performance. The next three experiments use the same basic setup to explore cases in which at least one of these assumptions is relaxed.

## 6. Experiment 2: Deterministic links, multiple root causes

Experiment 2 relaxes the root sparsity assumption and explores how well people learn networks from class DN. The difficulty order in Eq. (7) predicts that structure learning should be more difficult for class DN than for class DS, but we expected that people's inferences about class DN would still be relatively accurate.

### 6.1. Method

#### 6.1.1. Participants

29 members of the CMU community participated in exchange for course credit.

#### 6.1.2. Materials

Experiment 2 used the same general scenario described in

[5] Additionally, we report corroborating model comparisons based on log-likelihood values in Appendix C.

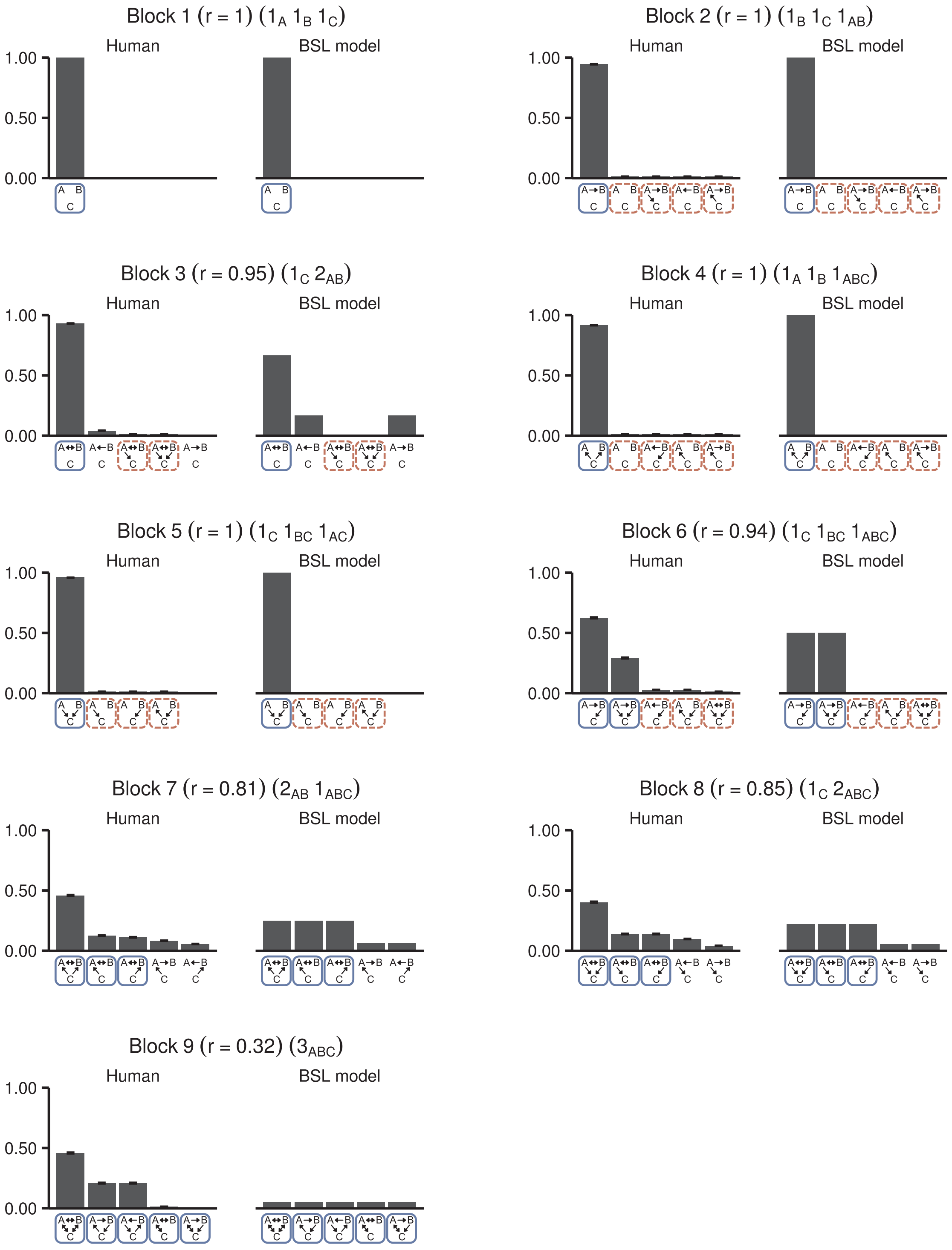


**Fig. 7.** Model predictions and human judgments for Experiment 1. Five out of the full set of 64 structures are included in each plot, and these five structures always include the two structures chosen most frequently by humans and the two most probable structures according to the model. Networks enclosed in solid blue boxes are the generating structures from Table 2. Networks enclosed in dashed red boxes are invalid responses that cannot explain at least one observation in a given block. Unboxed networks can account for each individual observation in a given block, but have characteristic distributions that do not match the distribution for the block. Error bars show standard errors based on bootstrap simulations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Experiment 1. The experiment included nine blocks that are shown in Table 3. Each three-node network in class DN now has 7 possible start states, which means that the characteristic set for each network includes 7 observations. The 64 three-node networks in class DN generate 9 characteristic distributions that are qualitatively different, and the blocks in Table 3 correspond to these nine characteristic distributions.

### 6.1.3. Procedure

The procedure was very similar to Experiment 1 with one key difference. During the introduction, participants were told that a single particle might directly activate one or more detectors. The three examples in the introduction included cases in which a single particle activated one, two, and four detectors respectively.

**Table 3**
9 Blocks used in Experiment 2.

| | Frequency | | | | | | | Generating structure |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC | |
| Block 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| Block 2 | - | 1 | 1 | 2 | - | 1 | 2 | |
| Block 3 | - | - | 1 | 3 | - | - | 3 | |
| Block 4 | 1 | 1 | - | 1 | - | - | 4 | |
| Block 5 | - | - | 1 | - | 2 | 2 | 2 | |
| Block 6 | - | - | 1 | - | - | 2 | 4 | |
| Block 7 | - | - | - | 3 | - | - | 4 | |
| Block 8 | - | - | 1 | - | - | - | 6 | |
| Block 9 | - | - | - | - | - | - | 7 | |

### 6.2. Results and discussion

The graded likelihood function for networks from class DN includes a background-rate parameter *b*. As suggested earlier, this parameter was set to the maximum likelihood estimate based on the examples participants observed in the introduction. In the introduction to Experiment 2, participants observed start states in which 1 out of 5, 2 out of 5 and 4 out of 5 nodes were active, suggesting that the background rate should be close to $\frac{7}{15} = 0.47$. This computation is not strictly correct because it does not acknowledge that start states must include at least one active node, and the actual maximum-likelihood estimate is $b = 0.44$. Additional details about the estimation procedure are provided in Appendix B, but the key point for now is that both models are fit without free parameters.

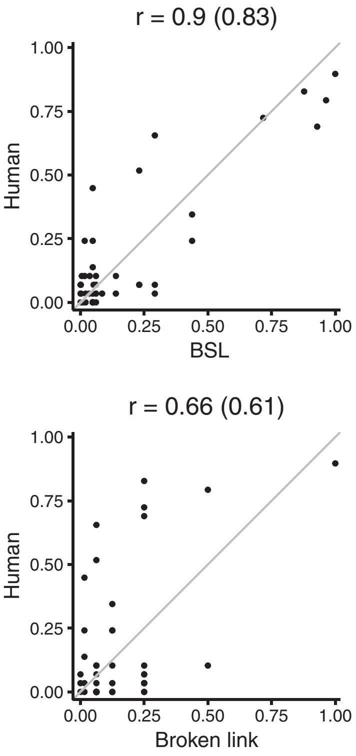


**Fig. 8.** Comparison of the complete set of human responses with model predictions for Experiment 2. In each panel the first correlation is based on the complete set of responses, and the correlation in parentheses shows the average correlation across the individual blocks of the experiment.

Fig. 8 summarizes the overall results. As for Experiment 1, the BSL accounts for the data relatively well, which indicates that people performed well at the task.

Results for the 9 individual blocks are shown in Fig. 9. As for Experiment 1, in every block the most common human response is correct. In particular, participants again reliably discovered the common effect structure in block 3 and the common cause structure in block 7. For blocks with multiple generating structures, the top few responses are often correct, but block 7 includes some cases in which a generating structure is less popular than a non-generating structure. These cases reveal some minor ways in which humans fall short of perfect performance on the task, but Fig. 9 suggests that the overall level of performance was high for each of the 9 blocks.

As for Experiment 1, comparing the BSL with the broken link model suggests that the graded likelihood plays an important role. As for Experiment 1, we ran a bootstrap analysis of the difference between the main correlations in Fig. 8. The difference of 0.24 between BSL and broken link is statistically significant, as the 95% confidence interval does not include zero, CI = [0.19, 0.28].

Allowing multiple root causes increases the number of structures that are consistent with a given data set, which means that the predictions of the broken link model become more diffuse. For example, the empty structure allows any observation to be explained by treating all active nodes as independent root causes, which means that the empty structure is now consistent with every block. For blocks 2 through 9, the broken link model cannot assign a probability to the generating structure that exceeds the probability assigned to the empty structure, which means that the generating structure can receive a probability of at most 0.5.

Comparing Experiments 1 and 2 suggests that causal determinism alone is sufficient to enable successful structure learning in our paradigm. It is therefore possible that root sparsity did not substantially contribute to our results for Experiment 1, but also possible that root sparsity is a second factor that is sufficient to enable successful structure learning. To adjudicate between these possibilities Experiment 3 explores structure learning when root sparsity holds but determinism does not.

## 7. Experiment 3: Probabilistic links, one root cause

Experiment 3 is directly analogous to Experiment 1 except that the determinism assumption is relaxed. Instead of assuming that activation always propagates along causal links, participants were

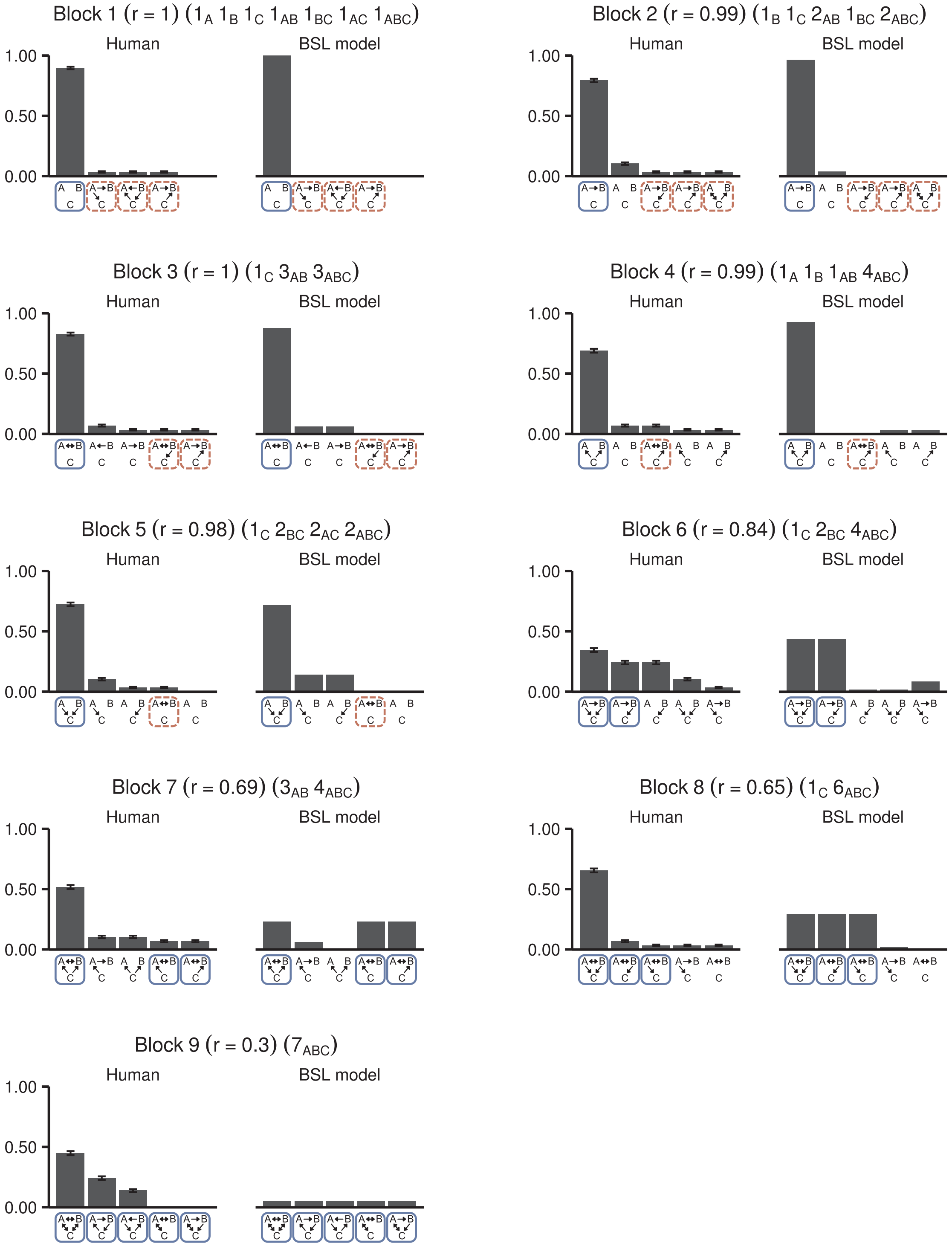


**Fig. 9.** Model predictions and human judgments for Experiment 2.

told that links in the underlying network were sometimes inactive and thus could fail to transmit activation. As for Experiment 1, participants were led to believe that root sparsity applied, which meant that at most one detector would spontaneously activate on any trial. The difficulty order in Eq. (7) predicts that structure learning is more difficult for class PS than for either class DS or DN, but it is still possible that the root-sparsity assumption alone will allow people to perform relatively well.

## 7.1. Method

### 7.1.1. Participants

29 members of the CMU community participated in exchange for course credit.

### 7.1.2. Materials

Experiment 3 used the same general scenario described in Experiment 1. In the instruction phase, active links between detectors were shown as green arrows and inactive links were shown as red

**Table 4**
12 Blocks used in Experiment 3.

| | Frequency | | | | | | | Generating structure |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC | |
| Block 1 | 4 | 4 | 4 | - | - | - | - | |
| Block 2 | 2 | 4 | 4 | 2 | - | - | - | |
| Block 3 | 2 | 2 | 4 | 4 | - | - | - | |
| Block 4 | 4 | 4 | 1 | - | 1 | 1 | 1 | |
| Block 5 | 2 | 2 | 4 | - | 2 | 2 | - | |
| Block 6 | 2 | 2 | 4 | 1 | - | 2 | 1 | |
| Block 7 | 2 | 2 | 2 | 4 | - | 1 | 1 | |
| Block 8 | 2 | 1 | 4 | 2 | - | 1 | 2 | |
| Block 9 | 1 | 1 | 4 | 1 | 1 | 1 | 3 | |
| Block 10 | 2 | 2 | 1 | - | 2 | 2 | 3 | |
| Block 11 | 2 | 2 | 2 | 1 | 1 | 1 | 3 | |
| Block 12 | 1 | 1 | 1 | 1 | 1 | 1 | 6 | |

arrows. Arrows drawn by participants were always black.

The experiment included 12 blocks that are shown in Table 4. Each three-node network in class PS has 3 possible start states and up to 64 graph variants, where the number of graph variants depends on the number of edges in the graph. Computing characteristic distributions for all networks produces 16 qualitatively different distributions, but some of these distributions cannot be represented using fewer than 12 observations. Table 4 includes all 12 distributions that can be represented using 12 or fewer observations. Some of these distributions can be represented using fewer than 12 observations—for example, the characteristic distribution corresponding to block 1 can be represented using three observations (one copy each of A, B and C). For consistency across the experiment, however, we used 12 observations for each block.

### 7.1.3. Procedure

The procedure was very similar to Experiment 1 except as noted below.

*7.1.3.1. Introduction*. During the introduction, participants were told that links between detectors could be active or inactive. Participants were told that every time a network was reset, the status of each link (active or inactive) was randomly determined. In the five introductory examples, active links were shown in green and inactive links were shown in red. No explicit information was provided about the probability of a link being active on a given trial, but the introductory examples were consistent with a base rate of 0.5 (13 active links out of 25).

*7.1.3.2. Observation phase*. During the observation phase of each block, participants were allowed to drag and sort observations in the observation panel. The blocks in Experiment 3 included more observations than the blocks in our previous experiments, and often included observations that were repeated several times. We hoped that allowing participants to sort the observations might make it easier for them to process each block.

*7.1.3.3. Test phase*. As for previous experiments, participants provided their responses by drawing black arrows between detectors. The instructions asked participants to report all links that existed regardless of whether they were active (green) or inactive (red) for a given observation.

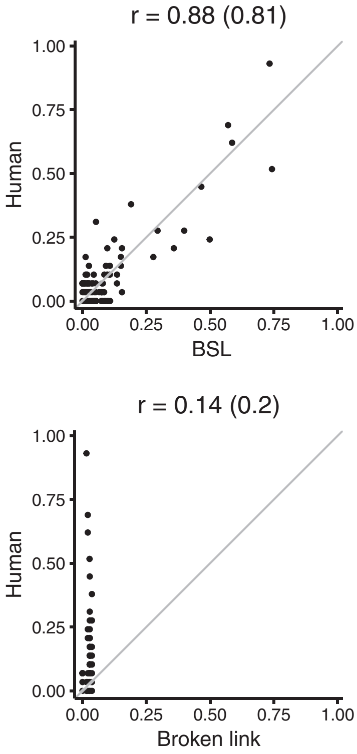


**Fig. 10.** Comparison of the complete set of human responses with model predictions for Experiment 3. In each panel the first correlation is based on the complete set of responses, and the correlation in parentheses shows the average correlation across the individual blocks of the experiment.

### 7.2. Results and discussion

The graded likelihood function for networks from class PS includes a failure rate parameter $f$. This parameter was set to $f = 0.48$ using maximum likelihood estimation based on the five introductory examples, each of which showed 2 or 3 out of 5 links as active and the other links as inactive (see Appendix B for details).

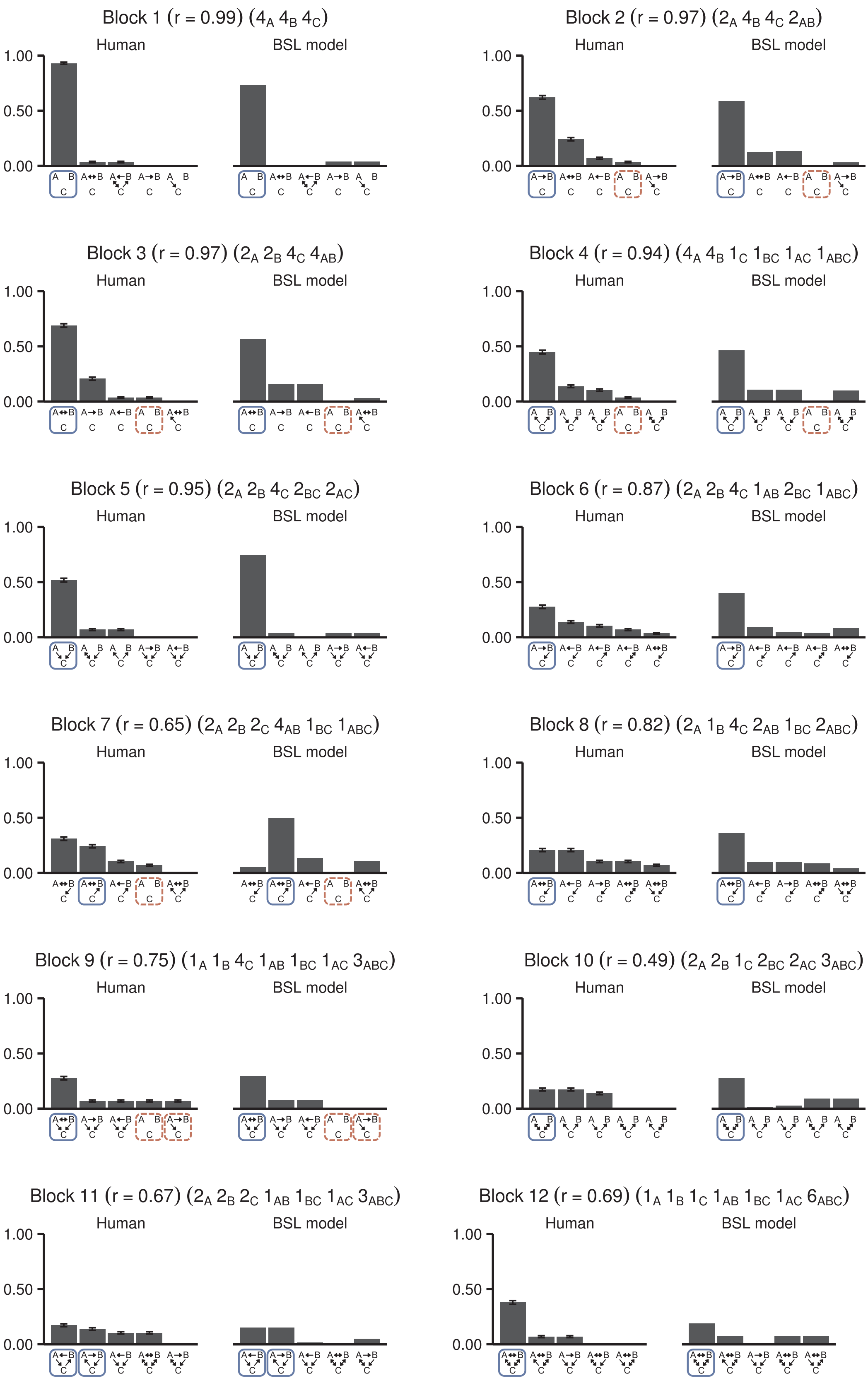


**Fig. 11.** Model predictions and human judgments for Experiment 3.

Fig. 10 summarizes the overall results. As for Experiments 1 and 2, the BSL captures the data relatively closely, which indicates that people performed well at the task.

Results for the 12 individual blocks are shown in Fig. 11. For 11 out of the 12 blocks, the most common human responses include at least one generating structure. In block 7, however, the most common response is not the generating structure. The generating structure is the second most popular response for this block, and differs from the most common response only with respect to the direction of the arrow between B and C. Even so, block 7 reveals a discrepancy between human reasoning and Bayesian inference that is greater than any discrepancy observed in Experiments 1 and 2. Overall, however, Fig. 11 suggests that the level of human performance was high for nearly all of the 12 blocks.

As for Experiments 1 and 2, comparing the BSL model with the broken link model suggests that the graded likelihood plays an important role. As before, a bootstrap analysis indicated that the difference in correlations between the two models shown in Fig. 10 was statistically significant, CI = $[0.70, 0.79]$.

Allowing probabilistic links dramatically increases the number of structures consistent with a given data set. For example, the fully connected structure is now consistent with every block, because appropriate patterns of edge failure allow this structure to account for every possible observation. As a result, the predictions of the broken link model become very diffuse, and the model cannot account for cases in which people are relatively confident about the true underlying structure.

Comparing Experiments 1 and 3 suggests that root sparsity alone is sufficient to enable successful structure learning in our paradigm. We can therefore conclude that root sparsity and determinism both enable successful structure learning. Performance when both assumptions hold is especially good (Experiment 1), but either assumption in isolation allows people to perform relatively well (Experiments 2 and 3). We now consider the final class of networks in Table 1, and explore structure-learning when neither assumption holds.

## 8. Experiment 4: Probabilistic links, multiple root causes

In three separate experiments we have found that people succeed at structure learning, and this finding contrasts with previous structure-learning studies that report relatively low levels of performance. These previous studies typically consider networks with probabilistic links and do not assume root sparsity. Experiment 4 follows suit and asks whether our experimental paradigm also leads to poor performance in the absence of determinism and root sparsity.

### 8.1. Method

#### 8.1.1. Participants

55 members of the CMU community participated in exchange for course credit.

#### 8.1.2. Materials

Experiment 4 used the same general scenario described in previous experiments.

In Experiments 1 and 2, the blocks were systematically constructed to include characteristic distributions that correspond to all possible three node networks. Using the same approach for Experiment 4 would produce 16 blocks, each of which includes 56 observations. An experiment using these blocks would be impractical, and we therefore simplified these 16 blocks in two ways.

First, given our expectation that learning the structure of PN networks would be difficult, we chose characteristic distributions corresponding to the 8 structures that seemed simplest and therefore easiest to learn. These structures are shown in Table 5, and include all structures with up to two directed links, and all structures for which all links are bi-directional.

Second, we approximated 5 of the 8 characteristic distributions rather than representing them exactly. Blocks 4–8 include between 12 and 14 observations each, and each block approximates a distribution that requires 28 observations to represent in full. We constructed these approximations in a way that aimed to maximize the statistical distinctions between blocks. The three remaining characteristic distributions (blocks 1–3) can each be represented using 14 observations each, which means that no approximation was required.

Of all four network classes considered in this paper, class PN comes closest to the class considered by most previous studies of structure learning. Even so, the notion of Markov equivalence does not apply to Experiment 4. For example, the network represented by Block 2 of Table 5 can be distinguished from a network with a single link from B to A even though the two networks are Markov equivalent. The networks have different characteristic distributions because causes are assumed to be generative, which means that nodes that receive links will be active more often than nodes that receive no links. For example, the distribution in Block 2 indicates that A will be active on 8 out of 12 trials, and B will be active on 10 out of 12 trials.

#### 8.1.3. Procedure

The procedure merged elements from Experiments 2 and 3. As for Experiment 2, participants were told that a single particle might directly activate one or more detectors. As for Experiment 3, participants were told that links between detectors could be active or inactive. The introduction included three introductory examples that supported these instructions. The observation and test phases for each block used the same procedure described for Experiment 3.

### 8.2. Results and discussion

As for experiments 2 and 3, the background-rate parameter $b$ and the failure-rate parameter $f$ were set using maximum-likelihood estimation based on the three introductory examples.

Fig. 12 summarizes the overall results. The correlation achieved by the BSL is substantially lower than for previous experiments (CI = $[0.39, 0.55]$, CI = $[0.32, 0.52]$, CI = $[0.29, 0.53]$, for the difference between the BSL correlation for Experiment 4 and the BSL correlations for Experiments 1, 2, and 3, respectively), suggesting that participants performed relatively poorly on the task. Fig. 13 shows results for the eight individual blocks, and reveals that participants often failed to recover the generating structure for each block.

Given the difficulty of the structure-learning task, it is important to consider whether the statistical evidence was sufficient to identify the generating structure for each block. For all of the 8 blocks, Fig. 13 confirms that the generating structure is indeed the structure with maximum posterior probability according to the BSL model. In some blocks, however, the difference in posterior probability between the generating structure and alternatives is arguably small enough that it may be unreasonable to expect participants to distinguish between these structures.

Although the task was intrinsically difficult, Fig. 13 nevertheless reveals some clear departures between human responses and Bayesian inference. Block 5 provides the clearest example. The model predictions for this block indicate that the available observations provided clear statistical support for the common effect structure ahead of the empty structure, but even so the common effect structure was chosen relatively rarely by participants. Overall, then, our data for Experiment 4 suggest that participants fell short of the benchmark provided by Bayesian inference.

As for previous experiments, the BSL model performs better than the broken link model. The binary likelihood no longer makes any contribution, because every structure is consistent with every possible observation if probabilistic links and multiple root causes are allowed. As a result the broken link model is completely unable to distinguish between structures.

Now that we have considered all four network classes in Table 1, we can conclude that determinism and root sparsity both enable successful structure learning, but that performance is relatively poor if neither assumption applies. Our final experiment asks whether one of these two assumptions is more fundamental than the other.

**Table 5**
8 Blocks used in Experiment 4. Asterisks indicate cases in which a block corresponds approximately but not exactly to the characteristic distribution of the generating structure.

| | Frequency | | | | | | | Generating structure | |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC | | |
| Block 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | | |
| Block 2 | 1 | 2 | 2 | 3 | 1 | 2 | 3 | | |
| Block 3 | 1 | 1 | 2 | 4 | 1 | 1 | 4 | | |
| Block 4 | 2 | 2 | - | 2 | 1 | 1 | 4 | | * |
| Block 5 | 1 | 1 | 2 | - | 3 | 3 | 4 | | * |
| Block 6 | 1 | 1 | 2 | 2 | 1 | 3 | 4 | | * |
| Block 7 | 1 | 1 | - | - | 2 | 2 | 7 | | * |
| Block 8 | - | - | - | 1 | 1 | 1 | 9 | | * |

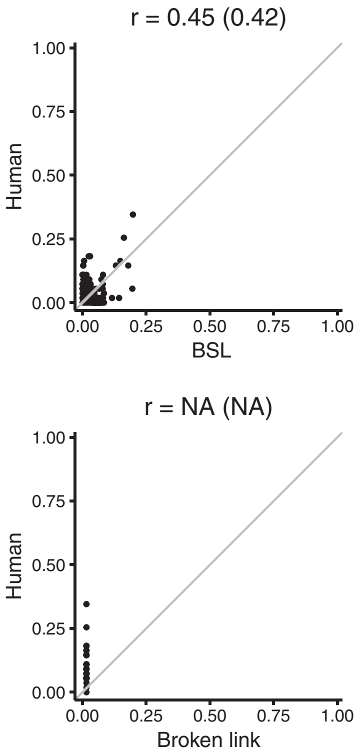


**Fig. 12.** Comparison of the complete set of human responses with model predictions for Experiment 4. In each panel the first correlation is based on the complete set of responses, and the correlation in parentheses shows the average correlation across the individual blocks of the experiment.

## 9. Experiment 5: Mostly deterministic links, mostly one root cause

Our results demonstrate that people are able to reason successfully about networks that violate the root sparsity assumption (Experiment 2) and networks that violate the determinism assumption (Experiment 3). Neither assumption is essential for structure learning to succeed, but it is possible that one assumption is psychologically privileged with respect to the other. Experiment 5 explored this possibility using a task that required participants to abandon either root sparsity or determinism, but not both. If one of these assumptions is privileged, than this privileged assumption should be preserved whenever possible, even at the cost of abandoning the other.

The idea of pitting determinism against root sparsity is related to the work of Mayrhofer and Waldmann (2016) on prior expectations in structure learning. Mayrhofer and Waldmann consider a structure-learning problem with two focal variables, one of which is the cause of the other. They describe two priors that could be used to infer the direction of the causal relationship. A *sufficiency* prior captures the idea that the causal relationship is very strong, and is related to the assumption of determinism that we have discussed throughout. A *necessity* prior captures the idea that the cause is necessary for the effect to occur, and is related to our notion of root sparsity.[6] Mayrhofer and Waldmann point out that these priors allow people to distinguish between Markov-equivalent structures, and consider a set of cases in which the two priors make different predictions. They find that some participants consistently match the predictions of the sufficiency prior, and that others consistently match the predictions of the necessity prior. Across three experiments, however, they find that participants match the sufficiency prior more closely than the necessity prior, suggesting that sufficiency may be psychologically more important than necessity. The results of Mayrhofer and Waldmann are consistent with a broader literature that suggests that people give more weight to sufficiency than necessity (Goldvarg & Johnson-Laird, 2001; Mandel & Lehman, 1998). Translating these findings into our terminology suggests that determinism may be psychologically more important than root sparsity.

One additional consideration suggests that determinism may be privileged with respect to root sparsity. Determinism appears to have received more attention than root sparsity in the psychological literature, suggesting that determinism may be the more fundamental concept. However, violating determinism (Experiment 3) did not seem to reduce people's level of performance more than did violating root sparsity (Experiment 2). The 95% confidence interval for the difference in correlations achieved by the BSL model across these two experiments is CI = [−0.07, 0.08], suggesting that the performance of the BSL model is comparable in both cases. As a result we had no strong expectation about the outcome of Experiment 5.

### 9.1. Method

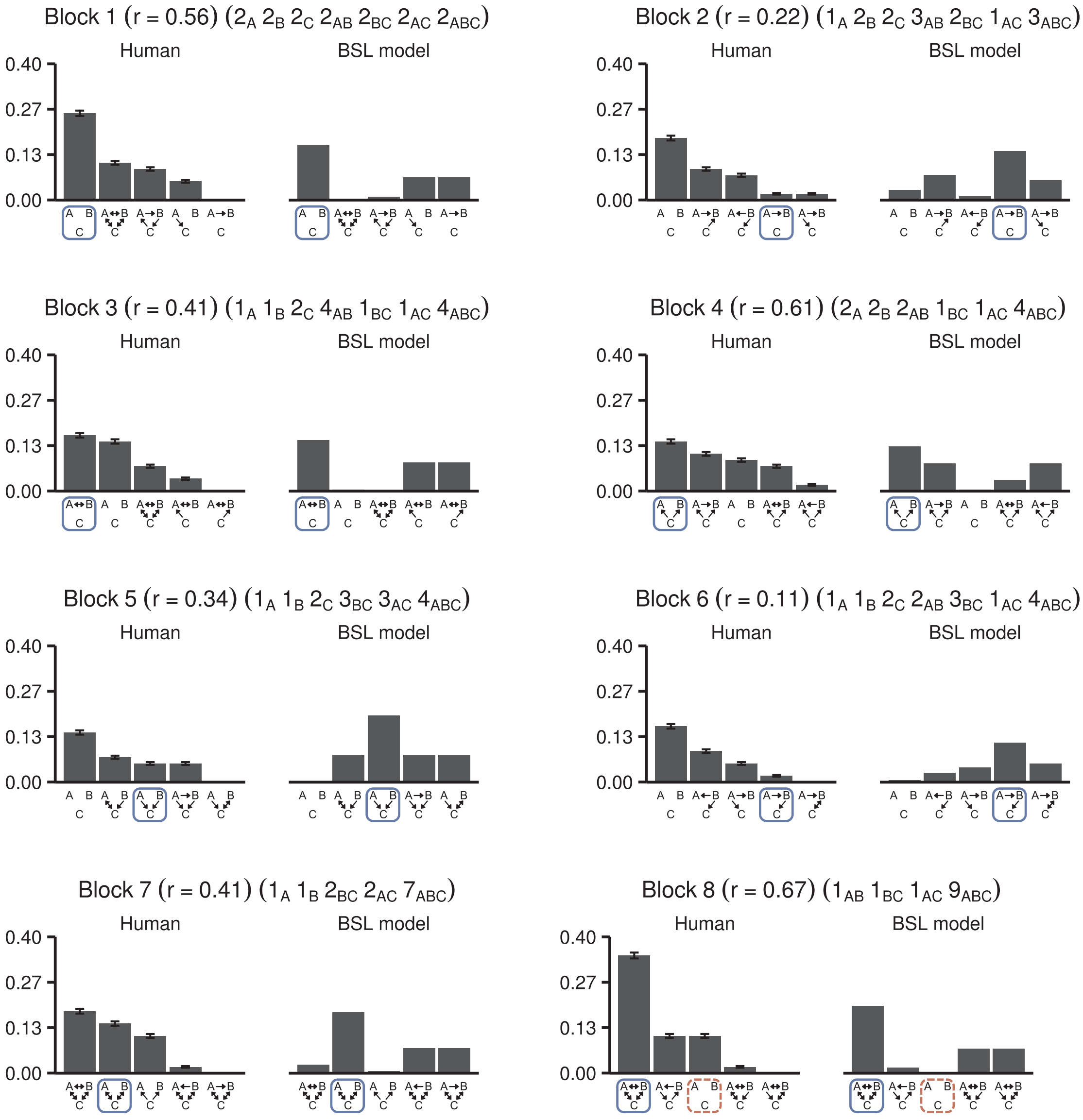
#### 9.1.1. Participants

24 members of the CMU community participated in exchange for course credit.

#### 9.1.2. Materials

Experiment 5 used the same general scenario described in previous

[6] Mayrhofer and Waldmann assume that causes act independently, and in the absence of interactions a necessary cause must be the only cause of an effect.


Block 1 (r = 0.56) ($2_A\ 2_B\ 2_C\ 2_{AB}\ 2_{BC}\ 2_{AC}\ 2_{ABC}$)

Block 2 (r = 0.22) ($1_A\ 2_B\ 2_C\ 3_{AB}\ 2_{BC}\ 1_{AC}\ 3_{ABC}$)

Block 3 (r = 0.41) ($1_A\ 1_B\ 2_C\ 4_{AB}\ 1_{BC}\ 1_{AC}\ 4_{ABC}$)

Block 4 (r = 0.61) ($2_A\ 2_B\ 2_{AB}\ 1_{BC}\ 1_{AC}\ 4_{ABC}$)

Block 5 (r = 0.34) ($1_A\ 1_B\ 2_C\ 3_{BC}\ 3_{AC}\ 4_{ABC}$)

Block 6 (r = 0.11) ($1_A\ 1_B\ 2_C\ 2_{AB}\ 3_{BC}\ 1_{AC}\ 4_{ABC}$)

Block 7 (r = 0.41) ($1_A\ 1_B\ 2_{BC}\ 2_{AC}\ 7_{ABC}$)

Block 8 (r = 0.67) ($1_{AB}\ 1_{BC}\ 1_{AC}\ 9_{ABC}$)

**Fig. 13.** Model predictions and human judgments for Experiment 4.

**Table 6**
10 Blocks of 3–4 observations used in Experiment 5. "Determinism" explanations can account for a block if links are always active and an additional root cause is invoked for one observation (all blocks except 5) or for two observations (block 5). "Root sparsity" explanations can account for a block if root causes are always sparse and an inactive link is invoked for one observation.

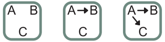
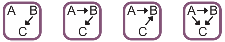
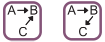
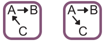
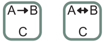
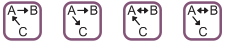
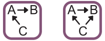
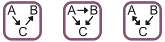
| | Frequency | | | | | | | Generating structure | |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC | Determinism | Root-sparsity |
| Block 1 | 1 | 1 | 1 | - | - | 1 | - | | |
| Block 2 | - | 1 | 1 | - | - | 1 | - |  |  |
| Block 3 | - | 1 | 1 | 1 | - | 1 | - | | |
| Block 4 | - | - | 1 | 1 | - | 1 | - | |  |
| Block 5 | - | - | 1 | 1 | - | 1 | 1 | | |
| Block 6 | - | 1 | 1 | 1 | - | - | 1 | |  |
| Block 7 | - | - | 1 | 1 | - | - | 1 |  |  |
| Block 8 | 1 | 1 | - | 1 | - | - | 1 | |  |
| Block 9 | - | - | - | - | 1 | 1 | 1 | |  |
| Block 10 | - | - | 1 | - | 1 | 1 | 1 | | |

**Table 7**
5 Filler blocks of 3 observations used in Experiment 5. Each filler block was presented twice.

| | Frequency | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | AB | BC | AC | ABC |
| Block 1 | 1 | 1 | 1 | – | – | – | – |
| Block 2 | 1 | 1 | – | – | – | 1 | – |
| Block 3 | 1 | 1 | – | – | – | – | 1 |
| Block 4 | 1 | – | – | 1 | 1 | – | – |
| Block 5 | 1 | – | – | 1 | – | – | 1 |

experiments, but the method used to construct blocks differed from that used in previous experiments. We considered only blocks that included at most one instance of each observation. Among these blocks, we used the 10 cases shown in Table 6 that are best explained by invoking either a minimal violation of determinism or a minimal violation of root sparsity.

To identify these 10 blocks, we first enumerated all blocks that include at most one instance of each observation. The shortest such block includes only one observation (e.g. {A}), and the longest includes seven ({A, B, C, AB, AC, BC, ABC}). We identified blocks that were equivalent up to relabeling, and retained a single representative of each equivalence class.

We then removed all blocks that could be generated over a DS network—in other words, all blocks that were consistent with both determinism and root sparsity. Of the blocks that remained, we selected those that could be explained by invoking either a single link failure or a single case in which two root causes were present. Table 6 includes explanations for each block that satisfy this condition.

Including only the 10 blocks used in Table 6 would mean that participants would need to invoke either a link failure or an extra root cause in every block, which would undermine the idea that violations of determinism and sparsity are rare. We therefore added 10 filler blocks that can be explained without invoking link failures or extra root causes. These blocks are shown in Table 7.

### 9.1.3. Procedure

The procedure was very similar to Experiment 4, except that link failures and extra root causes were described as possible but rare. The first introductory example was a case with a single root cause and no inactive links. The second example included a single link failure. Participants were told that “normally all connections are active after the network is reset”, but “on rare occasions all connections but one are active after the network is reset.” The third example included one extra root cause. Participants were told that “normally a particle directly activates only one of the detectors,” but “on rare occasions a particle simultaneously activates two of the detectors.” The presentation order of examples 2 and 3 was counterbalanced between participants.

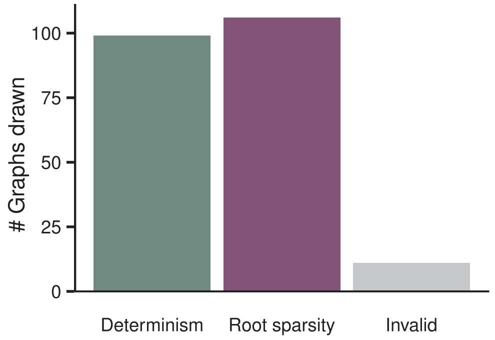


**Fig. 14.** Number of responses to Experiment 5 that preserved determinism, preserved root-sparsity, or were invalid. Counts are based on all blocks except block 5.

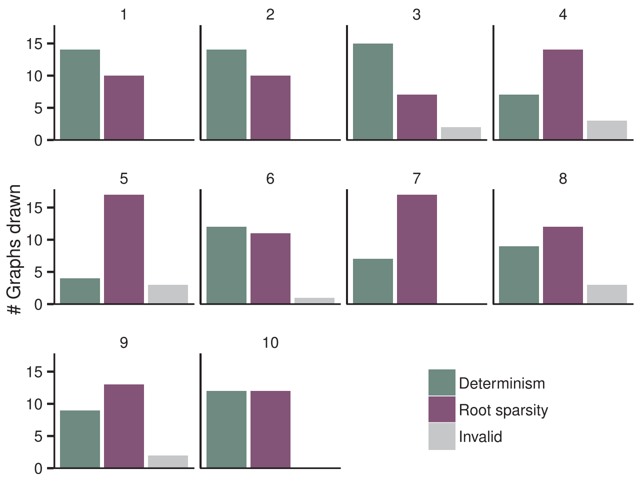


**Fig. 15.** Experiment 5 data summarized by blocks.

### 9.2. Results and discussion

Our analyses focus on the 10 critical blocks in Table 6, and the filler blocks will not be discussed. Fig. 14 summarizes how often participants preserved determinism and root sparsity when responding to the critical blocks. Responses are classified as determinism and root-sparsity responses based on whether they match the “determinism” and “root-sparsity” explanations in Table 6. Responses that do not match these explanations are classified as invalid, because they fail to provide parsimonious explanations of the data. Fig. 14 shows that determinism and root sparsity are preserved about equally often. If anything there is a slight preference for preserving root-sparsity rather than determinism.

Fig. 15 summarizes the responses for the ten individual blocks. Because block 5 is the only block that requires at least two violations of root-sparsity if determinism is preserved, it is perhaps not surprising that responses to this block tended to preserve root-sparsity rather than determinism. To enable a fair comparison, we therefore excluded block 5 when computing the summary results in Fig. 14. Block 7 produced a similar pattern of results (note that the observations for these blocks are very similar), but responses to the remaining blocks were not extremely skewed in favor of one assumption or the other.

At the population level, Figs. 14 and 15 suggest that neither determinism nor root sparsity is strongly privileged. Fig. 16 summarizes the responses for individual participants. We find that some participants consistently preserved determinism, others consistently preserved

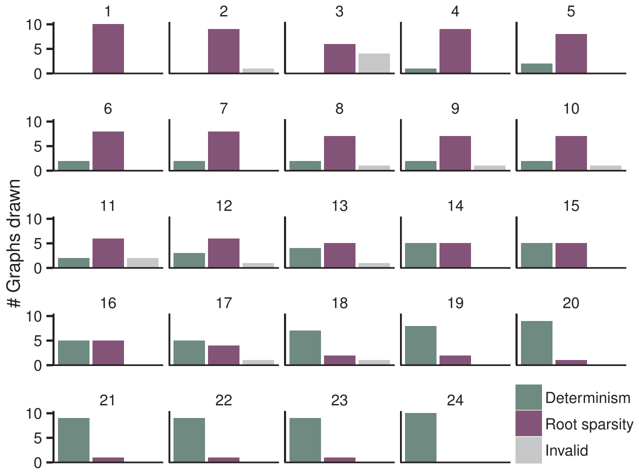


**Fig. 16.** Individual-level results for Experiment 5. Participants are ordered based on how consistently they preserve determinism. Counts are based on all blocks except block 5.

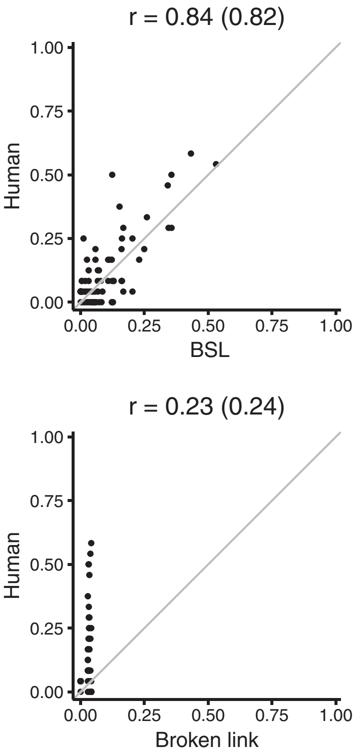


**Fig. 17.** Comparison of the complete set of human responses with model predictions for Experiment 5. In each panel the first correlation is based on the complete set of responses, and the correlation in parentheses shows the average correlation across the individual blocks of the experiment.

root-sparsity, but that there are also participants who preserved determinism on some trials and root-sparsity on others. This mixed pattern of responses is consistent with the individual differences reported by Mayrhofer and Waldmann (2016). Our data are therefore consistent with the view that some individuals view one assumption as more fundamental than the other, but also suggest that a substantial number of individuals show no such preference.

Taken together, the analyses in Figs. 14–16 provide little support for the hypothesis that determinism is privileged with respect to root-sparsity, or vice versa. It remains possible that one assumption is psychologically privileged, but that Experiment 5 was not sensitive enough to detect this difference. For example, the experimental instructions explicitly suggested that both determinism and root-sparsity could be violated on rare occasions, and perhaps these instructions overrode any natural bias that people have in favor of one assumption or the other. If such a bias exists, however, our results suggest that it must be fairly weak.

The primary result of Experiment 5 is summarized by Fig. 14, but for completeness our data are compared with model predictions in Fig. 17. The BSL provides a relatively good account of the data, suggesting that people succeed at structure learning even in cases in which violations of determinism and root-sparsity are both possible. The BSL model also accounts significantly better for the data than the broken link model, CI = [0.58, 0.65]. Results for individual blocks are shown in Fig. 18. For every block, the most common human response is among the best responses according to the BSL, suggesting that the model provides a relatively good account of the data for each individual block.

Experiments 4 and 5 are similar in that both allow for violations of determinism and violations of root-sparsity. These violations, however, are common in Experiment 4 but rare in Experiment 5. Comparing our data for the two experiments suggests that people perform relatively well at structure learning when violations of the two core principles are rare, but relatively poorly when violations of both principles are common.

## 10. Overall model comparison

Now that we have presented results from five separate experiments, we consider how well the BSL and the broken link models account for the complete set of data. The two models appear in the first two columns of Fig. 19, and the remaining columns are for alternative models that will be discussed in subsequent sections. The final row of Fig. 19 shows aggregate plots that summarize the performance of a given model across all five experiments. These aggregate results provide additional support for the conclusion that the BSL model performs better than the broken link model. The BSL model achieves a correlation of 0.89 across the complete set of data while the broken link model achieves a correlation of 0.64, and the difference is significant in a bootstrap analysis, CI = [0.22, 0.29]. Corroborating results based on log-likelihood values are reported in Appendix C.

## 11. Alternative models

Although the BSL model accounts well for many aspects of our data, at least two important questions remain to be addressed. The first concerns the role of the prior. We have relied on a uniform prior so far, but perhaps changing this prior would produce a Bayesian model that better captures people's inferences.

The second question asks how well the BSL model performs relative to non-Bayesian approaches that could be tried. Researchers have developed many formal accounts of causal learning from contingency data (Cheng, 1997; White, 2002), including models that focus on associative learning (Shanks & Dickinson, 1987), necessary and sufficient conditions (Mackie, 1965), and propositional reasoning (Boddez, Houwer, & Beckers, 2017). Most of these models focus on parameter learning, or learning the strength of the relationship between a candidate cause and an effect. These parameter-learning models, however, can be extended in order to handle structure-learning problems, and we will discuss the prospects of developing a successful model along these lines.

### 11.1. Bayesian models with alternative priors

The prior distribution $P(G)$ in Eq. (1) can capture tendencies that are not based on the data but that nevertheless lead people to prefer one graph over another. For example, people may have a tendency to prefer graphs that seem simple. Measures of graph simplicity have been developed by scientists from multiple fields (Mowshowitz & Dehmer, 2012), including computer science, biology, and chemistry, but there is little evidence to suggest which of these measures might have greatest psychological relevance. Here we consider two possible ways to formalize the notion of simplicity.

The first approach is to define the simplicity of a graph as a function of the number of edges that it contains. By this measure the empty graph is the simplest graph, and the fully connected graph is the most complex. Previous Bayesian approaches to structure learning sometimes use a simplicity measure of this kind to define a prior that favors graphs with few links (Dawid & Lauritzen, 1993; Jones et al., 2005). Fig. 21 shows a "few links" prior that sets the prior probability of a graph as inversely proportional to the number of links that it contains. The prior is defined over all 64 graphs with three nodes, but many of these graphs are equivalent up to relabeling of the nodes. Fig. 21 includes one representative from each of the 16 equivalence classes induced by node relabeling. Fig. 19 includes results for a "few links" model that combines the few links prior with the same likelihood function used by the BSL. The two models perform similarly, and the few links model does not emerge as superior to the BSL, CI = [−0.01, 0.01].

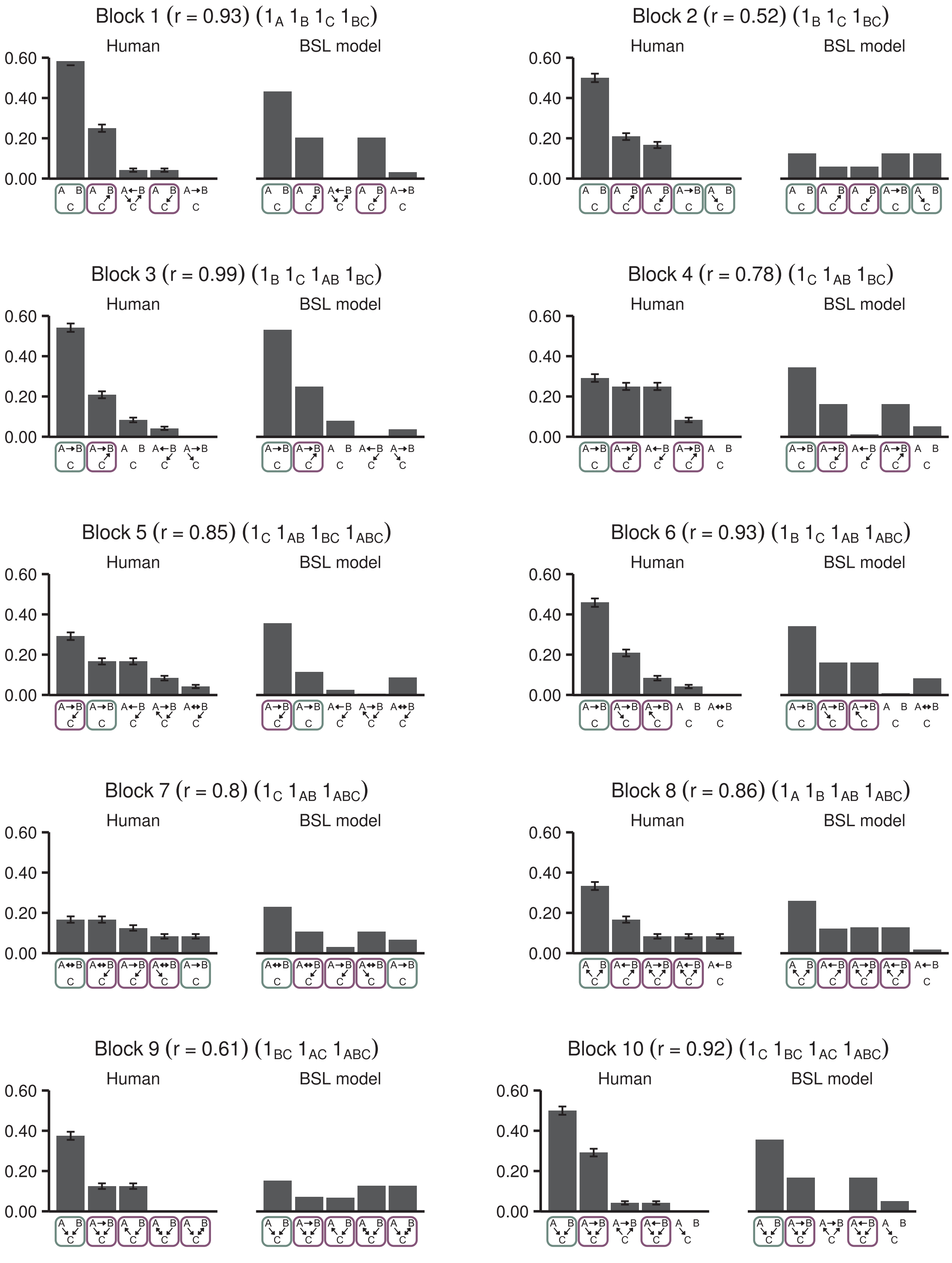


**Fig. 18.** Model predictions and human judgments for Experiment 5.

A second way to define the simplicity of a graph starts with the idea that simple graphs have many symmetries (Rashevsky, 1955). The symmetry score of a graph can be formally defined as the number of node permutations that leave the structure of the graph unchanged. There are six possible permutations of three nodes, including the "identity permutation" that maps each node to itself. The graph with no edges and the fully connected graph share the highest possible symmetry scores, because all six node permutations leave the structure of these graphs unchanged. The existence of the identity permutation means that the smallest possible symmetry score is 1. Fig. 21 includes a symmetry prior defined as

$$P(G) \propto s(G), \tag{8}$$

where $s(G)$ is the symmetry score of graph $G$.[7]

Fig. 21 includes results for a symmetry model that combines the

[7] Alternative ways to define a symmetry prior are possible. For example, Rashevsky proposes a symmetry-based measure of graph simplicity that leads to a prior that is different from the symmetry prior in Eq. (8). The two priors are closely related, but Eq. (8) is conceptually simpler than a prior based on Rashevsky's measure, which is defined using the notion of group orbits. For the hypothesis space of all 3-node structures, the correlation between the two priors is 0.92.

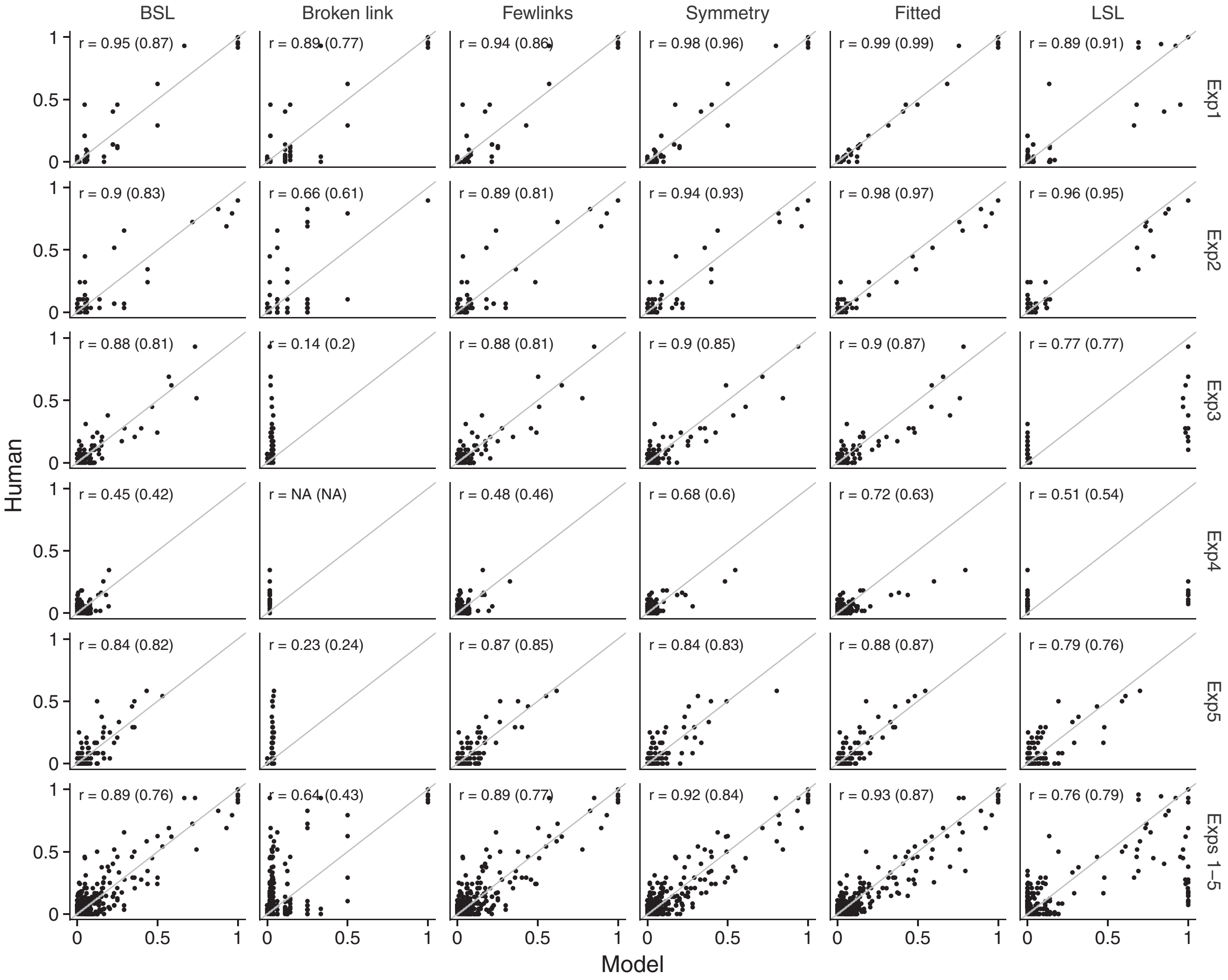


**Fig. 19.** Correlation plots for all models. The last row shows data collapsed across experiments.

symmetry prior with the same graded likelihood function used by the BSL. The symmetry model performs slightly better than the BSL overall, CI = [0.02, 0.04]. The differences between the models are most apparent when comparing average within-block correlations (e.g. 0.97 vs 0.87 for Experiment 1). These average correlations reveal that there are some blocks for which the symmetry model performs substantially better than the BSL. One example is Block 9 in Experiment 1 (see Fig. 7). After observing ABC three times in succession, the BSL assigns highest probability to the 18 structures that can only generate the observation ABC. These 18 structures correspond to all possible relabelings of the generating structures shown in Table 2 for block 9. Among these 18 structures, Fig. 7 shows that people overwhelmingly prefer the three graphs with highest symmetry scores: the fully connected graph and the two cycles. Other blocks that reveal a preference for symmetric structures include blocks 7 and 8 in Experiment 1, blocks 7, 8 and 9 in Experiment 2, block 12 in Experiment 3. Fig. 20 includes four of these blocks and shows in each case that the symmetry model accounts for human judgments better than the BSL.

Although the symmetry prior accounts well for our data, it is possible that this prior captures people's preferences only roughly. For example, perhaps people have an *a priori* preference for the empty graph and the fully connected graph, which happen to be the two graphs with highest prior probability according to the symmetry prior. The symmetry prior, however, makes additional distinctions between structures: for example, it assigns higher prior probability to common cause and common effect structures than to structures that include a single directed edge. It is possible that fine-grained distinctions like these are not reflected in people's judgments.

To address this possibility, we implemented a version of the Bayesian model for which the prior is fit to our data. Each prior distribution in Fig. 21 is characterized by 16 weights, one for each equivalence class of structures (groups of structures that are identical up to relabeling), and we allowed these weights to vary freely subject to the constraint that the resulting prior corresponded to a probability distribution over the hypothesis space of 64 graphs. The best-fitting weights are shown in Figs. 21, and 19 shows the performance of a model that combines this fitted prior with a graded likelihood.

The fitted prior in Fig. 21 does not correspond exactly to the symmetry prior. Note, for example, that the priors disagree with respect to the relative probabilities of the empty graph and the cycle. Most of the distinctions made by the symmetry prior, however, are also reflected in the fitted prior. For example, the symmetry prior distinguishes between six structures with a symmetry score of 2 (including common cause and common effect structures) and six structures with a symmetry score of 1 (including the structure with a single link), and this distinction is broadly consistent with the fitted prior.

Additional support for the symmetry prior is provided by Fig. 19. Comparing the columns for the symmetry and fitted prior models shows that replacing the symmetry prior with the fitted prior improves model performance by a small margin only, CI = [0.00, 0.02]. Overall, then,

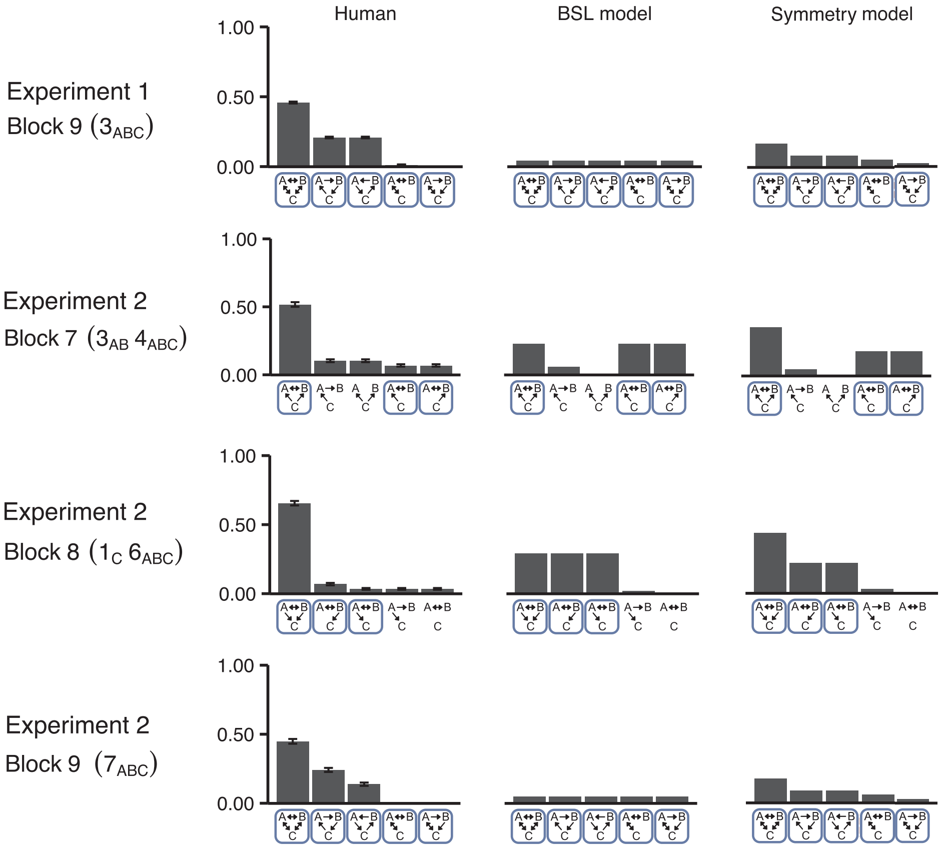


**Fig. 20.** Selected blocks in which the symmetry model captures qualitative distinctions that are missed by the BSL.

we can conclude that the distinctions captured by the symmetry prior are broadly consistent with our data, and that any additional distinctions made by people account for only relatively minor aspects of our data.

Our results for alternative priors suggests that the symmetry model is qualitatively superior to the BSL as a descriptive account of structure learning. The success of the symmetry model calls for further study of the reasons why the symmetry prior improves on the uniform prior used by the BSL. One possibility is that people believe that symmetric structures are more common than asymmetric structures, but other explanations are possible. Consider, for example, the three most common responses to block 7 in Experiment 1. Participants might understand that all three structures are equally compatible with the data, but feel that it is arbitrary to say that C sends a link to A but not B, or vice versa. Choosing the most symmetric among these three structures could be a way to signal that the observations provide no basis for breaking the symmetry between nodes A and B. Other response strategies are possible—for example, in cases where multiple structures are equally compatible with the data, participants might tend to choose the most symmetric structure simply because they have an aesthetic preference for symmetry. To us it seems likely that participants' responses reflect both prior beliefs about which structures are most probable and a range of different response strategies. From this perspective the symmetry prior makes accurate predictions about *how* people break ties between structures that are equally compatible with the data, but does not reveal the response strategies that are actually used to break these ties.

### *11.2. Process models of structure learning*

Like many Bayesian models, the symmetry model predicts how people respond to a task but does not characterize the psychological processes that actually generate their responses. It is therefore important to consider how the computations specified by the model could be implemented or approximated by psychological mechanisms.

One path towards a process model is to build on existing models of parameter learning such as the Rescorla-Wagner model. These models use contingency data to compute the strength of the relationship between a candidate cause and an effect. Because the candidate cause and the effect must be specified in advance, these models do not directly address the problem of structure learning. They can be extended in this direction, however, by carrying out parameter learning for each possible pair of variables, and assuming that all pairwise relationships greater than some threshold correspond to edges in a causal structure.

A pairwise approach should be able to successfully learn many of the structures used in our experiments, including the common effect structure that generated the data in Fig. 5. Combining pairwise inferences, however, is not enough to account for all aspects of our data. In Block 9 of Experiment 1, observations are generated over a fully connected graph, and any pairwise approach will assign the same strength to each pair of nodes. A pairwise model can therefore explain why the fully connected graph is the most common response to this block, but cannot explain why people prefer some subsets of this graph to others. For example, around half of the time participants chose one of the two 3-edge cycles, and a pairwise approach cannot explain why these cycles are chosen more often than other structures with three edges.

Although we suggest that no pairwise approach can provide a complete account of our data, we implemented one such approach in order to compare its performance with our Bayesian models. Because this model relies on local computations (Fernbach & Sloman, 2009; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008; Wellen & Danks, 2012) we refer to it as the "local structure learner", or the LSL for short.

Throughout the learning process, the LSL maintains a weight for

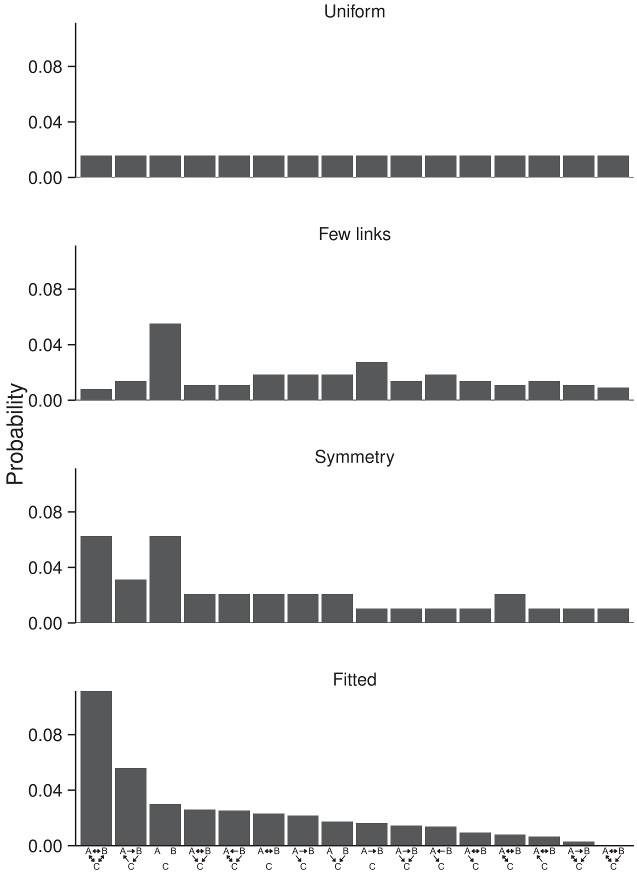


**Fig. 21.** Four possible priors over graphs.

each pair of nodes. Large positive weights indicate directed pairs that probably correspond to causal links, and large negative weights indicate directed pairs that probably do not correspond to links. For a problem involving three nodes, there are six weights in total, and each weight initially begins at zero. Suppose that the LSL now observes a state in which A and B are active but C is inactive. This observation supports the idea that there is a link from A → B, and the corresponding weight is updated by adding $p$. The weight for B → A is also incremented by the same amount. The same observation suggests that links A → C and B → C do not exist, and as a result both corresponding weights are decremented by subtracting $n$. The two remaining weights (corresponding to the links C → A and C → B) are left unchanged. The two parameters $p$ and $n$ both lie between 0 and 1, and we stipulate that $n + p = 1$. As a result, the weight update process relies on a single free parameter.

After all observations have been made, each weight is used to decide whether or not the corresponding causal link exists. Each weight $w$ is transformed into a choice probability using the function $(1 + e^{-sw})^{-1}$, where the slope $s$ is a second free parameter. These choice probabilities are then used to decide whether or not each edge exists. For example, a link that ends up with a weight of zero produces a choice probability of 0.5, and therefore has probability 0.5 of appearing in the final structure chosen by the model.

Fernbach and Sloman (2009) suggest that human causal learning is both structurally and temporally local, and the LSL is local in both senses. The model is structurally local because it relies on computations that are carried out separately for each pair of nodes. The model is temporally local because it does not require all observations to be stored for subsequent batch processing. Instead, each observation can be forgotten after it is used to update the current set of weights.

To give the LSL the best possible chance of performing well, we fit the free parameters $p$ and $s$ to each experiment separately, and the resulting parameter values are shown in Table 8. It is natural to think that people might focus more on positive evidence (i.e. evidence for the existence of a link) than negative evidence (i.e. evidence that a link does not exist). If the positive evidence parameter $p$ is set to 1, then the LSL reduces to a model that considers only positive evidence. This special case of the LSL, however, performs poorly, and when the $p$ parameter is allowed to vary freely, Table 8 shows that the best-fitting parameter values tend to assign greater weight to negative evidence than positive evidence.[8]

**Table 8**
LSL parameters that maximize the correlation between model predictions and human responses to each experiment.

| Experiment | $p$ | $s$ |
|---|---|---|
| 1 | 0.11 | 14.2 |
| 2 | 0.06 | 8.2 |
| 3 | 0.53 | 37.6 |
| 4 | 0.00 | 15.3 |
| 5 | 0.26 | 4.1 |

Fig. 19 summarizes the performance of the LSL. Compared to the symmetry model, the LSL performs slightly worse for Experiments 1 and 5, and slightly better for Experiment 2. For Experiments 3 and 4, however, the LSL performs substantially worse than the symmetry model, and these experiments mean that the aggregate performance of the LSL (r = 0.76) is worse than both the symmetry model (r = 0.92), CI = [0.13, 0.19], and the BSL model (r = 0.89), CI = [0.10, 0.17]. There is therefore a significant difference in favor of the probabilistic models even though the LSL has two free parameters per experiment and the symmetry and BSL models both have none.

For Experiment 3, the LSL ends up with a very large slope parameter, which means that it assigns non-negligible probability to only a single structure per block. The model therefore does not capture the graded nature of people's responses that is evident in blocks 2, 5, and 6 of Fig. 11. An additional qualitative failure of the LSL emerges directly from the local nature of the model. The LSL makes essentially the same inference about blocks 11 and 12, and assigns very high probability to the fully-connected structure in both cases. People, however, tend to choose the two 3-edge cycles in block 11. Between them, the two cycles include all 6 possible edges, and considering each edge in isolation means that the LSL will either accept all of them, none of them, or some random subset of them depending on the value of the slope parameter. Regardless of how the model parameters are set, the model has no way of preferring structures with certain holistic properties – for example, structures in which the edges are arranged into a cycle.

For Experiment 4, the LSL makes essentially the same inference for every block, and assigns very high probability to the empty structure and negligible probability to the remaining structures. In contrast, Fig. 19 suggests that the symmetry model is able to capture some of the distinctions that people make in Experiment 4.

Despite the qualitative problems just identified, the model achieves a number of successes. For example, even though Experiment 3 exposes some problems with the model, the best response according to the LSL is the most common (or equal most common) human response for 10 out of the 12 blocks in this experiment. There may be ways to adjust the LSL to allow it to perform even better. Even so, the analyses in this section tend to suggest that the symmetry model captures aspects of people's inferences that go beyond simple local computations.

The LSL is an especially simple model, and other structure learning models that emphasize local computation may provide a better account of our data. Bramley, Dayan, Griffiths, and Lagnado (2017) describe an online structure learning algorithm that maintains a single candidate

[8] See Table C.1 for model parameters that maximize the likelihood of the data.

structure at any stage and updates it by adjusting one edge at a time. Like the LSL, this incremental approach seems unlikely to capture people's preferences for structures with holistic properties such as symmetry. In principle, however, it should be possible to develop process models that do take symmetry into account.

One possible approach is to develop a hybrid approach that combines the LSL (or the approach of Bramley et al. (2017)) with the symmetry model. For example, the LSL could be used to generate a handful of candidate structures for a given problem, and these candidates could be subsequently evaluated using the symmetry model to identify the candidate with maximum posterior probability. A hybrid approach along these lines could help to explain how the computations required by the symmetry model are implemented (or approximated) in a way that is psychologically plausible.

## 12. Discussion

We presented five experiments that explore structure learning from observational data. Our studies focused specifically on two assumptions: determinism and root sparsity. We found that both assumptions enabled successful structure learning (Experiments 1 through 3). If violations of both assumptions are common then performance is relatively poor (Experiment 4), but performance remains high if violations are possible but rare (Experiment 5).

The most pressing question raised by our results is why people performed relatively well in our experiments but relatively poorly in previous structure-learning experiments. We have touched on this question already but now take it up in some detail, and discuss previous work on learning both deterministic and probabilistic causal structures.

### 12.1. Learning deterministic structures

As mentioned earlier, our first experiment is closely related to White's work on learning the structure of deterministic causal systems (White, 2006). In most of White's experiments, the unobserved causal structure is a food web in which the nodes represent animal species and the edges represent predator-prey relationships. Participants were told about changes in the populations of the species. For example, one observation might specify that only the population of species B changed during a given season. A second might specify that the populations of species A and B both changed during another season. These two observations can be summarized using the observation set {B, AB}, and this set suggests the existence of a causal link between A and B (compare with our opening example in Fig. 1).

In contrast to our results, White found that participants were mostly unable to recover the correct structure for a given set of observations. Performance remained poor even when White gave his participants explicit instructions about how to infer the underlying causal structure. White's studies were specifically designed to explore cases in which temporal information and co-occurrence information provide competing cues to causal structure, and it is therefore not surprising that we were able to find conditions under which performance exceeds the levels reported by White. Even so, it is useful to consider some of the specific factors that may contribute to the difference between the two sets of results.

First, assumptions about background causes are critical for solving both White's task and ours, but these assumptions may not have been clear to White's participants. For example, White's first experiment included two structure-learning problems, and the observation set for one of these problems was {A, B, C, D, ABCDE}. If the underlying network is a DS network, then the network must include exactly four links, one from E to each of the other species. White, however, did not appear to communicate the assumption of root-sparsity to his participants, or the related assumption that background causes are rare. If there is no particular reason to think that background causes are rare, then observation ABCDE can be explained by invoking five separate root causes, and the set of five observations does not provide strong support for the "correct" causal structure that White had in mind. Unlike White, we explicitly conveyed the notion of root-sparsity to our participants. For example, the introduction to our first experiment included three examples that were designed in part to illustrate this notion.

A second characteristic of White's experiments is that his materials seem unintuitive in several respects. One issue is that the directionality of causal links in a food web is somewhat ambiguous. White's participants were supposed to think that links were directed from predator species to prey species, but it seems just as natural to think that changes in prey populations can cause changes in predator populations.

A third issue is that White deliberately described observations such as {A, B, C, D, ABCDE} as observations that were gathered in five successive seasons. White reports that many of his participants were sensitive to the temporal order of the observations, and focusing on temporal information may have prevented them from approaching the problem using the normative method that White outlines. In our experiments, we made it clear that the networks were "reset" between observations, eliminating any temporal interpretations.

There are other differences between White's experiments and our own that might partially explain the difference between our respective findings. White focused on two causal structures, each of which had 5 nodes, but our experiments focused on structures with 3 nodes each. In comparing his work with previous work by Gopnik, Sobel, Shulz, and Glymour (2001), White suggests that 3-node problems may be small enough for people to compute normative solutions, but that for larger problems computing normative solutions "imposes too great a demand on working memory, or is for some other reason too difficult to accomplish" (p. 476). We suspect, however, that White's 5-node problems can be readily solved in the context of our particle-detector paradigm. Some preliminary support for this claim is provided by a post-test given to our Experiment 1 participants after the experiment proper had ended. The post-test included two structure-learning problems based on the same 5-node structures and observation sets used by White, and we found that participants were reliably able to recover the correct structures (see Deverett & Kemp (2012) for additional details). This post-test does not provide strong evidence that our materials are more intuitive than White's, because the difference between experimental materials is confounded with a difference in whether or not the 5-node problems followed a test involving 3-node problems. The post-test does demonstrate, however, that there are conditions under which people are capable of learning the 5-node structures used by White.

### 12.2. Learning probabilistic structures

Although White considered deterministic causal systems, other recent psychological work on structure learning has focused on probabilistic causal systems. The standard finding is that systems with probabilistic links are difficult to learn. We found that root-sparsity allowed participants to infer the structure of probabilistic systems (Experiment 3), but that probabilistic systems were difficult to learn in the absence of root-sparsity (Experiment 4). The primary question raised by our data is how our positive results for Experiment 3 can be reconciled with previous negative results for probabilistic systems. We focus in particular on the previous work of Steyvers et al. (2003) and Lagnado and Sloman (2004).

Steyvers et al. (2003) asked participants to learn the structure of 3-node networks. Nodes in each network corresponded to aliens, and links between nodes represented cases in which an alien was able to read another alien's mind. The values of the nodes represented words in the aliens' minds, and each node was categorical with a large number of possible values. In their first experiment, Steyvers et al. (2003) focused on learning from observational data, and made the task especially simple by giving participants a forced choice between two possible structures—a common effect structure and a common-cause structure.

Even so, performance on the task was relatively poor, and around half of the in-lab participants (other participants completed the task over the web) performed at chance.

Unlike the systems used in Experiment 3, the alien mind-reading networks did not satisfy the assumption of root-sparsity. For example, on a typical trial, aliens A and C might both be thinking the word POR and alien B might be thinking the word TUS. If A is reading C's mind (or vice versa), then a single root cause can explain the data for A and C, but an additional root cause is needed to explain why B is thinking TUS instead of one of many other possible words. Based on our results, we predict that performance on the alien task would improve if the paradigm were adjusted to satisfy the assumption of root-sparsity. For example, if the aliens are allowed to have empty minds, then each observation could involve a single word that is present in the minds of one or more aliens. The resulting task would be closely related to our task in Experiment 3, and we expect that participants would find it relatively straightforward.

Lagnado and Sloman (2004) carried out a second set of structure-learning experiments that focus on 3-node networks. Their cover stories were based on two realistic scenarios: for example, in the chemist scenario the true causal structure was a chain in which acid level (low or high) influenced ester level (low or high), which in turn influenced whether or not perfume was produced. Both causal links in the chain were probabilistic with a causal strength of 0.8. In their first experiment, Lagnado and Sloman asked participants to infer which of 5 causal structures generated a set of observations. Performance was poor, and only 14% of participants selected the correct structure.

Unlike the alien mind-reading networks used by Steyvers et al., the causal chain used by Lagnado and Sloman did satisfy the assumption of root sparsity. In fact, this causal chain allowed only one possible root cause—the initial node in the chain—because there were no background causes that could activate the second or third node if the first variable in the chain was inactive. The causal chain task is therefore directly related to our third experiment, which asked people to learn the structure of probabilistic networks from class PS. Our data for block 5 of this experiment show that our participants were able to reliably reconstruct a probabilistic causal chain over three variables, which contrasts with the result of Lagnado and Sloman.

One possible explanation for the difference is that our participants were given information about the functional forms of the causal relationships in the underlying network, but Lagnado and Sloman's participants were given no such information. Expectations about functional form appear to be critical for generating the correct response to Lagnado and Sloman's causal chain task. In terms of our notation, the observations available for this task correspond to a distribution over the set {ABC, AB, A}. Different assumptions about functional form lead to different inferences about the underlying structure. If links are probabilistic but only the root variable in a causal structure can be spontaneously active, then the data support the correct solution according to Lagnado and Sloman (i.e. a chain with links from A to B and B to C). If links are deterministic and spontaneous activations are possible for all nodes, then the data support a causal chain in the opposite direction (i.e. with links from C to B and from B to A). In block 5 of Experiment 1, most of our participants made exactly this inference, and did so in part because they knew that the underlying network belonged to class DS.

We have highlighted one way in which our setup differs from the Steyvers et al. mind-reading task, and a second way in which it differs from the Lagnado and Sloman causal chain task. Other differences between these tasks, however, deserve some attention. There is an important respect in which our task was more difficult than both previous tasks. We asked participants to draw a graph over three nodes, which effectively requires them to choose one among 64 possible structures. The two previous studies were much more constrained, and required participants to choose either one among two structures (Experiment 1 of Steyvers et al.) or one among five structures (Experiment 1 of Lagnado and Sloman). From this perspective, the high level of performance observed across our experiments is all the more striking.

There are other respects, however, in which our task was easier than both previous tasks. We minimized demands on memory by leaving all observations on screen at the time when participants made their inferences, but Steyvers et al. and Lagnado and Sloman both presented at most one observation at a time. We did not include empty observations (i.e. observations for which all detectors were inactive), but Lagnado and Sloman did, and half of the 50 observations that they showed fell into this category. The frequency of empty observations is informative about base rates but uninformative about causal structure, and dropping these observations may have helped our participants to focus on the information that is most relevant to structure learning. Finally, we believe that our "activation detector" scenario is more intuitive than the scenarios used by both previous studies. One issue with the mind-reading task is that the direction of causal links clashes with intuitions about causal agency (Mayrhofer & Waldmann, 2015). For example, if A is reading B's mind, A is the active partner and therefore naturally viewed as the cause of the interaction between the two, but the link in the causal network over these nodes is directed from B to A. One concern with the causal-chain task is that both real-world scenarios used by Lagnado and Sloman may have led to the expectation that the underlying structure should be a common effect structure rather than a chain.

We have now identified several possible factors that help to explain why people performed better in our experiments than in previous studies of structure-learning. There may be other relevant factors, and additional experiments are needed to determine which differences between the paradigms in question are truly critical. We hope, however, that the differences singled out in this section help to make the discrepancy between our results and previous results less puzzling than it may initially seem.

### 12.3. Bayesian vs constraint-based approaches to structure learning

Computer scientists have developed two prominent approaches to structure-learning: the constraint-based approach (Pearl, 2000; Spirtes et al., 2001) and the Bayesian approach (Friedman & Koller, 2002; Heckerman, Meek, & Cooper, 1999). The constraint-based approach runs standard statistical tests to identify dependence or independence relationships between subsets of variables, and uses the outcomes of these tests to infer the underlying causal structure. The Bayesian approach incorporates a likelihood function that captures how the data were generated over an underlying causal structure, and computes a posterior distribution that combines this likelihood function with a prior over structures.

Both of these approaches have influenced psychological work on structure learning (Gopnik et al., 2004; Steyvers et al., 2003), but our work is more consistent with the Bayesian approach than the constraint-based approach. The BSL and symmetry models are instances of the Bayesian approach, and demonstrate that this approach accounts well for our data. The symmetry model incorporates a graded likelihood term and a symmetry-based prior, and we found support for both components of the model. The likelihood term is consistent with the finding that people tend to prefer structures that make the observed data not only possible but likely. For some blocks in our experiments, there were multiple structures that maximized the likelihood term, and among these maximum likelihood structures we found that people tended to prefer those that were symmetric. This result supports the Bayesian view that people bring prior expectations to structure learning.

There are several reasons to believe that a constraint-based approach would account less well for our data. First, some of our experiments used very small observation sets—for example, each block in Experiment 1 included three observations only. Constraint-based approaches rely on standard statistical tests, and these tests are not appropriate when the number of available observations is small. Second, constraint-based approaches do not naturally incorporate the kinds of prior beliefs that people appeared to bring to our task. These beliefs include expectations about causal structure, such as the expectation that symmetric structures are more likely than asymmetric structures. They also include expectations about functional form, such as the expectation in Experiment 1 that causal links were deterministic and that root causes were relatively rare.

In literature influenced by the constraint-based approach, it is common to find discussions of Markov equivalence classes, and the idea that networks from the same Markov equivalence class cannot be distinguished using statistical data. For example, the common-cause structure with links from A to B and A to C belongs to the same equivalence class as the chain with links from B to A and A to C, and these two structures are indistinguishable in the absence of expectations about functional form. In Experiment 1, however, we found that people readily distinguish between common-cause and chain structures (blocks 4 and 5), and the reason is that they were able to exploit expectations about functional form. Expectations about determinism and root sparsity place especially strong constraints on structure learning, but many other kinds of expectations are enough to render Markov equivalence irrelevant. For example, even the weak expectation that causes tend to be generative rather than preventive is enough to allow Markov equivalent networks to be distinguished. The notion of Markov equivalence is important in settings where prior knowledge is minimal, but we believe that human structure learning typically requires prior expectations of various kinds, and that Markov equivalence rarely poses a problem in everyday learning settings.

### 12.4. Human and model performance across the four network classes

Our experiments were organized around four classes of causal networks and our main result is that people performed well when reasoning about networks from classes DS, DN, and PS. This result contrasts with previous work which tends to suggest that people are relatively poor at structure learning.

Beyond the basic result that people sometimes succeed at structure learning, our data support comparisons across the four classes of networks. As suggested earlier, probabilistic non-sparse systems (PN) are intrinsically more complex than the other kinds of systems, and it is therefore not surprising that PN emerges as the most difficult class in our experiments. A more revealing finding is that all of the models match human judgments better for class DS than for class PN. A complete theory of human structure learning should produce excellent model fits across all of our experiments, including experiments where people are near-normative (Experiment 1) and experiments where they are not (Experiment 4). None of our models meets this goal, and future work should aim to develop a single theory that accounts well for human inferences across a wide variety of network classes.

### 12.5. From Bayesian networks to functional causal models

Causal structure learning is often formulated as the problem of learning a causal Bayesian network, and this approach has been productive. An alternative, however, is to formulate structure learning as the problem of learning a functional causal model. Fig. 4 illustrates how the activation networks considered in this paper can be represented as functional causal models. The key step is to introduce exogenous variables (such as $U_A$ and $U_{AB}$ in Fig. 4) so that each of the original variables is a deterministic function of its parents.

An important advantage of working with functional causal models is that these models can accommodate feedback loops and other kinds of causal cycles. Our experiments considered activation networks that can include cycles, and these networks are better captured by functional models than by causal Bayesian networks. Pearl (2000) offers some additional reasons why functional causal models should be preferred to Bayesian networks: for example, functional models better support inferences about counterfactuals (Lucas & Kemp, 2015) and actual causation (Halpern & Hitchcock, 2010).

Because functional models are more general than Bayesian networks, learning a functional model should be more difficult than learning a Bayesian network if generic priors are used in both cases. As suggested earlier, however, generic priors rarely seem appropriate for the learning problems that people face. In addition to exploring how people respond to generic structure learning problems, an important research direction is to explore how people learn functional causal models given different kinds of prior beliefs. Several kinds of prior beliefs are relevant to the work in this paper, including beliefs about generative causes, disjunctive causal combinations, determinism, and sparsity. Future work can explore structure learning in settings that draw on other kinds of prior beliefs, including settings in which causes may be preventive and in which conjunctive causal interactions are expected.

### 12.6. Other cues to causal structure

Because previous studies often found that structure learning is difficult given observational data alone, there has been considerable interest in other kinds of information that facilitate structure learning. In particular, multiple researchers have found that temporal information (Lagnado & Sloman, 2004; Rottman & Keil, 2012) and the ability to perform interventions (Lagnado & Sloman, 2004; Steyvers et al., 2003) both lead to improved performance.

Our result that observational data can be sufficient for structure learning provides a basis for future studies that explore how observational data relates to other sources of structural information. Previous studies have typically explored this question using tasks in which people fail at structure learning given observational data alone. Our paradigm could be adapted to explore different cues to structure in a setting in which observational data alone enables reliable inferences. For example, future studies could explore how heavily observational data is weighted relative to other sources of information, and how people resolve conflicts when there are multiple incompatible cues to causal structure.

### 12.7. Determinism and root sparsity

Our work focused on the implications of determinism and root sparsity for causal structure learning, but these factors are also relevant to other aspects of causal reasoning. For example, determinism suggests that people tend to generate explanations of an event that make the event seem inevitable, and root sparsity suggests that people often invoke a single root cause when explaining an event. Beyond causal reasoning, determinism and root sparsity are also relevant to other aspects of cognition, such as categorization. For example, determinism suggests that categories have crisp definitions that support unambiguous judgments about category membership, and root sparsity is consistent with the view that all of the properties of a category stem from a single underlying essence.

Previous work brings out the relevance of determinism, root sparsity or both to causal explanation (Chi et al., 2012; Lombrozo, 2007; Zemla et al., 2017), categorization (Gelman, 2003), decision-making (Gaissmaier & Schooler, 2008), and social attribution (Heider, 1958). Our work is consistent with the general themes emerging from many of these studies, but suggests a different perspective on determinism and root sparsity than is typically presented. Determinism and root sparsity are often viewed as cognitive biases that lead to faulty inferences and suboptimal decisions. In contrast, our results suggest that determinism and root sparsity can facilitate inferences (such as discovering the causal structure of a system) that would otherwise be difficult or impossible. Future work should consider how broadly this favorable view of determinism and root sparsity is likely to extend. A starting point is to consider cases in which these expectations are consistent with the structure of the environment, and therefore likely to be helpful rather than harmful. For example, root sparsity may be especially relevant when observing events in continuous time, because it is extremely unlikely that two or more unrelated causes would trigger at the same instant. Determinism may be helpful when reasoning about systems with hidden causal variables, because aspiring towards a deterministic explanation may lead people to discover latent variables that would otherwise have been overlooked (Schulz & Sommerville, 2006).

Although determinism is sometimes consistent with the structure of the environment, people seem to act as if this assumption applies more broadly than it actually does. For example, after an unexpected event occurs, people often show “creeping determinism” and report that it was predictable all along (Fischhoff, 1975). Similarly, the literature on probability matching suggests that participants often attempt to find deterministic patterns in sequences that are generated by random processes (Gaissmaier & Schooler, 2008). There is also some evidence that people tend to rely on root sparsity more than they should. For example, Lombrozo (2007) found that participants preferred to explain an outcome by invoking a single root cause, even when a two-cause explanation was actually more probable. On the other hand, Zemla et al. (2017) found that participants rated explanations with more root causes as more plausible than explanations with fewer root causes.

An important direction for future work is to characterize and explain the different expectations about determinism and root sparsity that people bring to different learning problems. For example, Yeung and Griffiths (2015) suggest that expectations about determinism are stronger for physical systems than for social systems. Similarly, Johnson et al. (2017) suggest that root sparsity is more likely to be assumed when reasoning about physical systems than when reasoning about social systems.

Expectations like these may be formed by generalizing over previous experience with physical and social systems. For example, after exposure to a number of different physical systems, a learner may notice that apparently unpredictable outcomes can usually be explained as a deterministic function of variables that were not initially obvious. As a result, the learner may assume that novel physical systems are also likely to be intrinsically deterministic. This kind of learning requires interaction with individual causal systems (e.g. operating a specific CD player) and experience with physical systems in general. Learning at both of these levels can be formalized using a hierarchical Bayesian approach. Hierarchical Bayesian models of causal learning have previously been explored by several groups of researchers (Hagmayer & Mayrhofer, 2013; Kemp, Goodman, & Tenenbaum, 2010; Lucas & Griffiths, 2010), and the same basic approach may be able to learn which assumptions about determinism and root sparsity are appropriate for a given class of problems.

## 13. Conclusion

Previous studies often report that structure learning from observational data is difficult. In contrast, our results suggest that people find structure learning relatively easy if causes are deterministic or if each observation has a single root cause. There may be additional factors that enable successful structure learning, and our work may lead to future efforts that comprehensively chart the conditions under which people perform well at structure learning from observational data.

Comparing our results with previous results suggests the need for psychological models that can explain why people perform well on some structure-learning tasks but poorly on others. We found that a Bayesian approach accounted well for our data, and exploring psychologically-plausible implementations of this approach may ultimately lead to a model that captures both people’s successes and their failures at structure learning tasks.

## Declarations of interest

None.

## Acknowledgments

Experiment 1 was presented at the 34th Annual Meeting of the Cognitive Science society. Experiments 2, 3, and 5 were written up as AR’s Master’s thesis completed at Georg-August University Göttingen under the supervision of RM and CK.

We thank Evan Metsky for help with running Experiments 3 and 4, and Alan Jern, York Hagmayer, Bob Rehder, and Michael Waldmann for suggestions about this work.

BD was supported by a training program in Neural Computation that was organized by CMU’s Center for the Neural Basis of Cognition and funded by NIH R90 DA023426. RM was supported by research grants of the Deutsche Forschungsgemeinschaft (DFG-Wa 621/22-1, DFG-Ma 6545/1-2) as part of the priority program “New frameworks of rationality” (SPP 1516). CK was supported in part by the James S. McDonnell Foundation Causal Learning Collaborative Initiative and by NSF award CDI-0835797.

## Appendix A. Supplementary material

The data for Experiments 1–5 can be found at https://osf.io/rx8fa/.

## Appendix B. Parameter settings

For each experiment, model parameters $b$ (see Eq. (4)) and $f$ (see Eq. (6)) were set to maximum likelihood estimates based on the observations shown in the instructions (Table B.2). Importantly, these parameters were based on observations that were *shown* to participants, not on data that were *collected* from participants. Consequently, these parameter values were fixed throughout the data analysis and not treated as free parameters.

The estimate of $f$ is equal to the relative frequency of broken links in the observations. For example, for Experiment 4 the estimate is $f = 0.48$ because 12 of 25 links were broken. As mentioned in the text, the estimate of $b$ must allow for the fact that each observed pattern of activation is the result of at least one root cause. For example, for Experiment 4 the estimate is $b = 0.44$, which maximizes the joint probability of the three observations with 1, 2, and 4 root causes respectively:

$$\underset{b}{\operatorname{argmax}} \left( \frac{\binom{5}{1} b^1 (1-b)^4}{1-(1-b)^5} \right) * \left( \frac{\binom{5}{2} b^2 (1-b)^3}{1-(1-b)^5} \right) * \left( \frac{\binom{5}{4} b^4 (1-b)^1}{1-(1-b)^5} \right)$$

Table B.3 shows all parameter values. As intended, these estimates were close to either 0 or 0.5 for the first four experiments. For completeness, Table B.3 also lists the number of root causes and broken links, which are relevant for the binary likelihood of the broken link model. These numbers come directly from the observations shown in the instructions (see Table B.2).

**Table B.1**
Five blocks with 1 or 2 observations and nine blocks with 6 observations used in Experiment 1.

| | Frequency | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | AB | AC | BC | ABC |
| Block 1 | 1 | – | – | – | – | – | – |
| Block 2 | – | – | – | 1 | – | – | – |
| Block 3 | – | – | – | 2 | – | – | – |
| Block 4 | – | – | – | – | – | – | 1 |
| Block 5 | – | – | – | – | – | – | 2 |
| Block 6 | 2 | 2 | 2 | – | – | – | – |
| Block 7 | – | 2 | 2 | 2 | – | – | – |
| Block 8 | – | – | 2 | 4 | – | – | – |
| Block 9 | 2 | 2 | – | – | – | – | 2 |
| Block 10 | – | – | 2 | – | 2 | 2 | – |
| Block 11 | – | – | 2 | – | – | 2 | 2 |
| Block 12 | – | – | – | 4 | – | – | 2 |
| Block 13 | – | – | 2 | – | – | – | 4 |
| Block 14 | – | – | – | – | – | – | 6 |

**Table B.2**
Frequencies and colors of observations presented during the instruction phase. Experiment 1 used red to indicate active detectors. Experiment 2 used red for some participants and green for others, and no difference in response patterns was found between these groups. Experiments 3–5 used green for active detectors to match the color used for active links.

| | Active links | Initially active detectors (root causes) | Link color (active/inactive) | Detector color (active/inactive) |
|---|---|---|---|---|
| Experiment 1 | 5 out of 5 | 1 out of 5 | Black/– | Red/gray |
| | 5 out of 5 | 1 out of 5 | Black/– | Red/gray |
| | 5 out of 5 | 1 out of 5 | Black/– | Red/gray |
| Experiment 2 | 5 out of 5 | 1 out of 5 | Black/– | Red/gray — green/gray |
| | 5 out of 5 | 2 out of 5 | Black/– | Red/gray — green/gray |
| | 5 out of 5 | 3 out of 5 | Black/– | Red/gray — green/gray |
| Experiment 3 | 3 out of 5 | – | Green/red | Green/gray |
| | 2 out of 5 | – | Green/red | Green/gray |
| | 3 out of 5 | 1 out of 5 | Green/red | Green/gray |
| | 3 out of 5 | 1 out of 5 | Green/red | Green/gray |
| | 2 out of 5 | 1 out of 5 | Green/red | Green/gray |
| Experiment 4 | 3 out of 5 | – | Green/red | Green/gray |
| | 2 out of 5 | – | Green/red | Green/gray |
| | 3 out of 5 | 1 out of 5 | Green/red | Green/gray |
| | 3 out of 5 | 2 out of 5 | Green/red | Green/gray |
| | 2 out of 5 | 3 out of 5 | Green/red | Green/gray |
| Experiment 5 | 5 out of 5 | 1 out of 5 | Green/red | Green/gray |
| | 4 out of 5 | 1 out of 5 | Green/red | Green/gray |
| | 5 out of 5 | 2 out of 5 | Green/red | Green/gray |

*Note.* Experiment 3 and 4 began with two "dry" examples that did not show active detectors yet and only introduced the idea of active and inactive links.

**Table B.3**
Parameter settings for the graded and binary likelihoods. The parameters for the binary likelihood represent the maximum number of root causes and broken links allowed for each experiment.

| | Graded likelihood | | Binary likelihood | |
|---|---|---|---|---|
| | $b$ | $f$ | # Root causes | # Broken links |
| Experiment 1 | 0.000001 | 0 | 1 | 0 |
| Experiment 2 | 0.44 | 0 | 3 | 0 |
| Experiment 3 | 0.000001 | 0.48 | 1 | 6 |
| Experiment 4 | 0.44 | 0.48 | 3 | 6 |
| Experiment 5 | 0.14 | 0.07 | 2 | 1 |

## Appendix C. Model comparison using log-likelihoods

The model comparisons in the main text rely on correlations. To check whether the resulting conclusions about the relative performance of the models are robust, we carried out a second set of comparisons using log-likelihood as the evaluation measure. For increased rigor, we used the log-likelihoods of out-of-sample predictions. As for the analysis in Fig. 18, we evaluated the models separately for each experiment.

The log likelihood of model $M$ for a given experiment is

**Table C.1**
LSL parameters that maximize log likelihood for each experiment.

| Experiment | $p$ | $s$ |
|---|---|---|
| 1 | 0.13 | 4.86 |
| 2 | 0.07 | 2.73 |
| 3 | 0.46 | 0.84 |
| 4 | 0.28 | 0.32 |
| 5 | 0.26 | 2.54 |

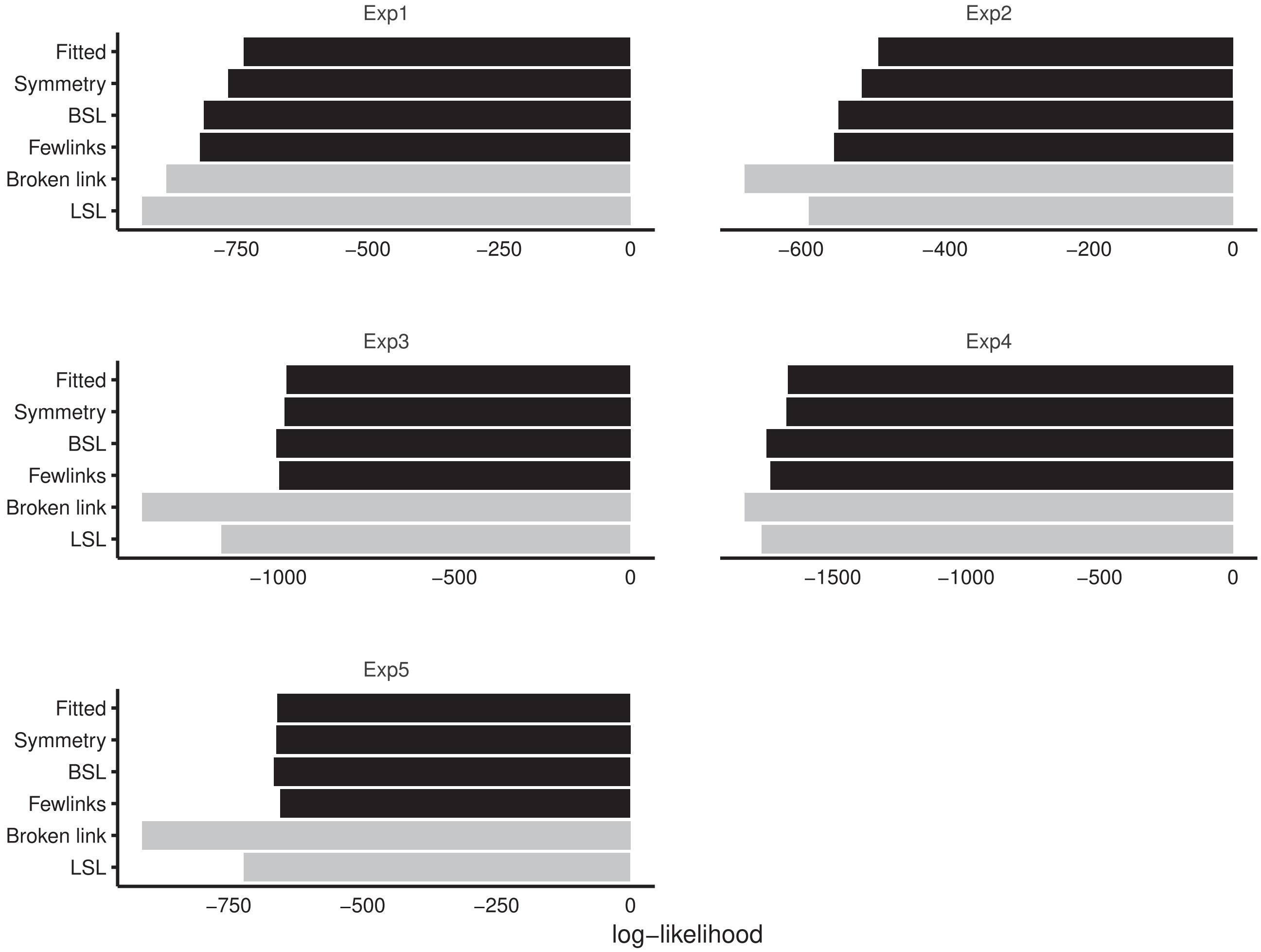


**Fig. C.1.** Model performances based on log-likelihoods. Larger log-likelihoods (i.e. likelihoods closer to 0) indicate better predictions about participants' behavior. All predictions were out-of-sample predictions. The broken link and the LSL model were outperformed by the other models, supporting the findings based on correlations reported in the main text. For visual guidance, the models are ordered by their performance in Experiment 1 and the two process models (broken link and LSL) are colored in gray.

$$\log P^*(D|M) = \sum_{i=1}^{I} \log P^*(d_i|M) = \sum_{i=1}^{I} \sum_{b=1}^{B} P^*(d_{ib}|M) \tag{C.1}$$

where $D$ is the full set of responses for the experiment, $I$ is the number of participants, $d_i$ is the full set of responses of participant $i$, $B$ is the number of blocks in the experiment, and $d_{ib}$ is the response of participant $i$ for block $b$.

Because some of the models assign zero probability to some responses, we supplement each model with a guessing parameter $\theta$ for every participant. We assume that response $d_{ib}$ is made randomly with probability $\theta_i$, and is drawn from the model's posterior with probability $1-\theta_i$:

$$P^*(d_{ib}|M, \theta_i) \propto (1-\theta_i)P(d_{ib}|M) + \theta_i P(d_{ib}|M_0) \tag{C.2}$$

where $P(\cdot)$ is the original model posterior that does not allow for guessing, $P^*(\cdot)$ is the posterior that allows for guessing, and model $M_0$ assigns uniform probability to all structures, $p(d_{ib}|M_0) = \frac{1}{64}$. For simplicity, we use a uniform prior on $\theta_i$ and integrate over all values:

$$P(d_{ib}|M) = \int_0^1 P(d_{ib}|M, \theta_i)p(\theta_i)d\theta_i \tag{C.3}$$

For each participant $i$, we approximated the integral using a discrete grid on $\theta_i$ with step size of 0.05.

Out-of-sample predictions provide a sensible comparison of models with different numbers of free parameters. For each participant, we fitted the models with free parameters to the remaining participants (optimizing for likelihood instead of correlation) and then computed for every model the log-likelihood of the data from this held-out participant.

The two models with free parameters were the model with the fitted prior (15 free parameters rather than 16 because the 16 weights plotted in Fig. 21 must sum to 1) and the LSL, which has two free parameters. The parameters of both models were fit to maximize the likelihood in Eq. (C.1). To keep the fitting procedure tractable we did not integrate over a guessing parameter but maximized the log-likelihood based on the original model posterior (i.e. we maximized $\log P$ instead of $\log P^*$). To evaluate the impact of maximizing for log-likelihood instead of correlation, we fitted the prior for the fitted model based on all participant data. This new fitted prior was almost identical to the prior shown in Fig. 21, and the correlation between the two was greater than 0.999. The parameters of the LSL that maximize the log-likelihood of the full set of participant data are shown in Table C.1.

The results based on out-of-sample log-likelihoods confirmed the results based on correlations in the main text. Fig. C.1 shows the out-of-sample log-likelihoods for each experiment. The Bayesian structure learner (BSL) had a higher likelihood than the broken link and local structure learner (LSL) models. Replacing the uniform prior used by the BSL with a symmetry prior improved the likelihood still further. The highest likelihood overall was achieved by the fitted model, suggesting that the fitted model outperforms the symmetry model even when making out-of-sample predictions. This result, however, may not be theoretically informative. Even though the fitted model accounts well for the data, it is hard to say why the fitted prior in Fig. 21 is psychologically more natural than the symmetry prior. In our view, the most important message of Fig. C.1 and of the correlation-based analyses in the main text is that the BSL and its two refinements (symmetry and fitted models) outperform the broken link and LSL models.

## Appendix D. Additional figures

Figs. D.1 and D.2

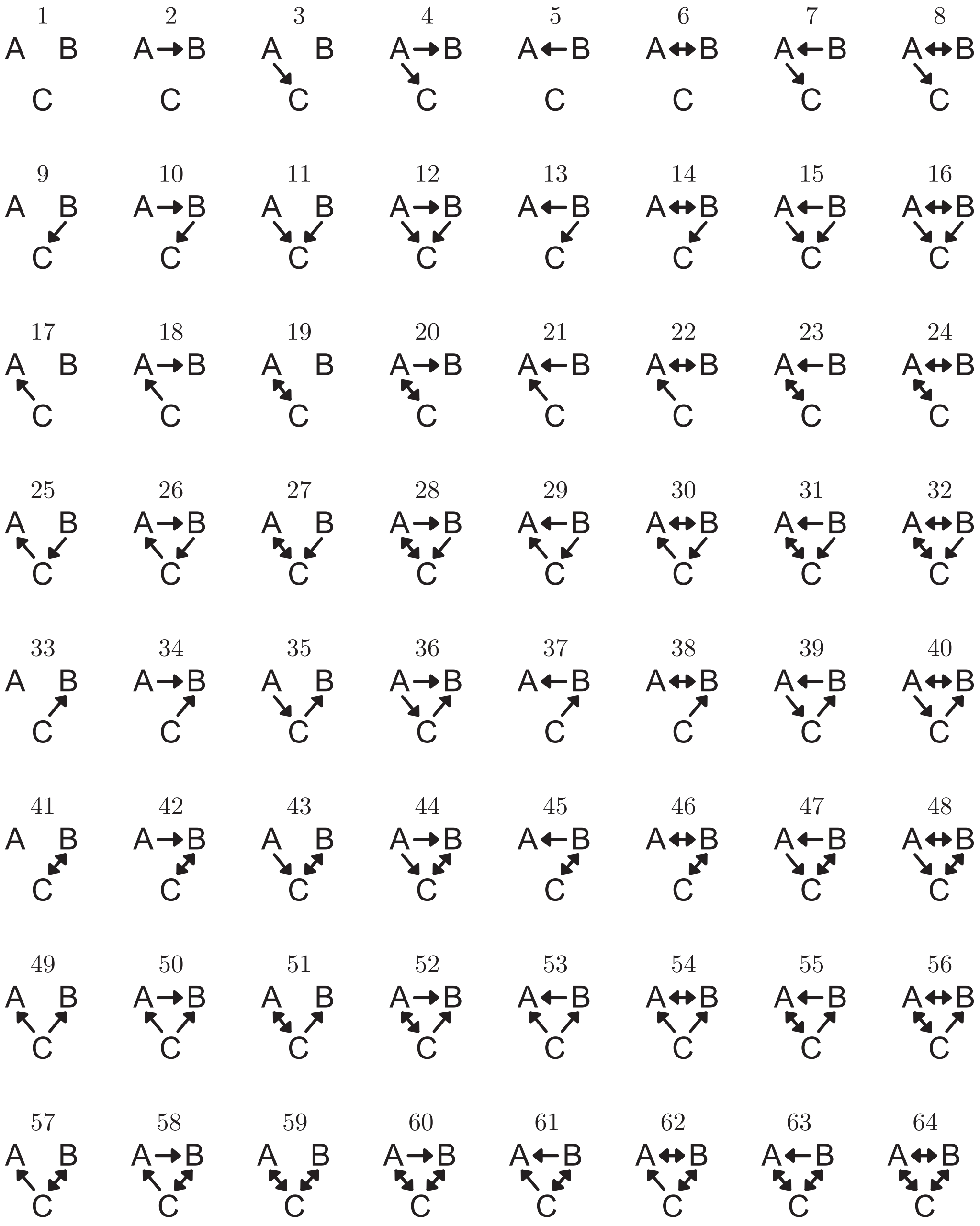


**Fig. D.1.** All possible directed graphs over three nodes.

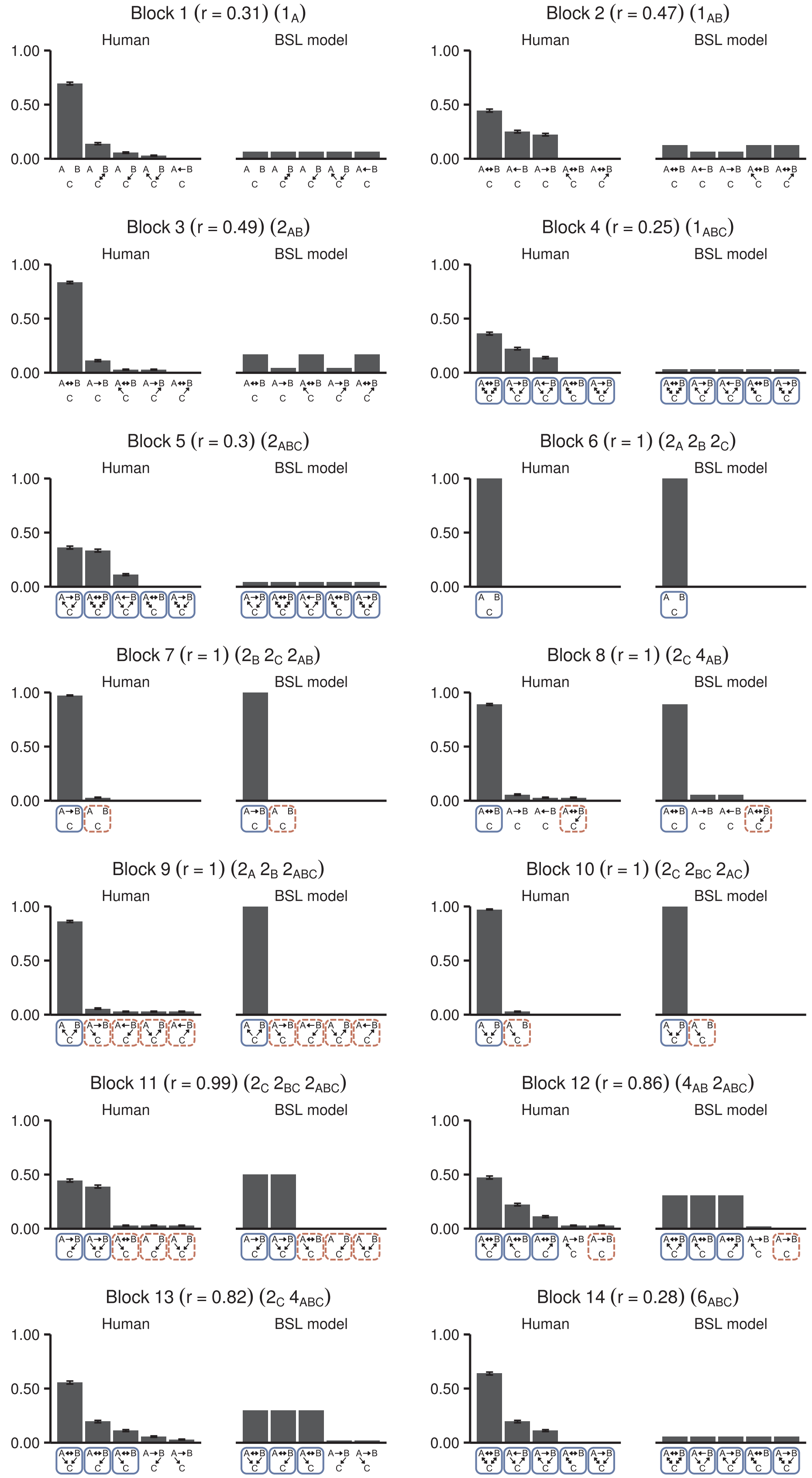


**Fig. D.2.** Model predictions and human judgments for additional blocks in Experiment 1.

## References

Boddez, Y., Houwer, J. D., & Beckers, T. (2017). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. *The Oxford handbook of causal reasoning* (pp. 53). Berlin: Oxford University Press.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48*, 1156–1164.

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

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

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*, 1–61.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models.* MIT Press.

Dawid, A. P., & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics,* 1272–1317.

Deverett, B., & Kemp, C. (2012). Learning deterministic causal networks from observational data. *Proceedings of the 34th annual conference of the cognitive science society* (pp. 288–293). Austin, TX: Cognitive Science Society.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines, 21*, 389–410.

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

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance, 1*, 288.

Friedman, N., & Koller, D. (2002). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning, 50*, 19–125.

Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica, 137*, 280–291.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition, 109*, 416–422.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* Oxford: Oxford University Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565–610.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1–31.

Gopnik, A., Sobel, D. M., Shulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*, 620–629.

Hagmayer, Y., & Mayrhofer, R. (2013). Hierarchical Bayesian models as formal models of causal reasoning. *Argument & Computation, 4*, 36–45.

Halpern, J. Y., & Hitchcock, C. R. (2010). Actual causation and the art of modelling. In H. Geffner, & J. H. R. Dechter (Eds.). *Heuristics, probability, and causality: A tribute to Judea Pearl* (pp. 383–406). London: College Publications.

Heckerman, D., Meek, C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In C. Glymour, & G. F. Cooper (Eds.). *Computation, causation and discovery* (pp. 141–166). Cambridge, MA: MIT Press.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*, 8–21.

Johnson, S. G. B., Valenti, J. J., & Keil, F. C. (2017). Opponent uses of simplicity and complexity in causal explanation. *Proceedings of the 39th annual conference of the cognitive science society* (pp. 606–611). Austin, TX: Cognitive Science Society.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., & West, M. (2005). Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science, 20*, 388–400.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science, 34*, 1185–1243.

Kim, N. S., Luhmann, C. C., Pierce, M. L., & Ryan, M. M. (2009). The conceptual centrality of causal cycles. *Memory & Cognition, 37*, 744–758.

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

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*, 232–257.

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

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review, 122*, 700–734.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955–984.

Mackie, J. L. (1965). Causes and conditions. *American philosophical quarterly, 2*, 245–264.

Mandel, D. R., & Lehman, D. R. (1998). Integration of Contingency Information in Judgments of Cause, Covariation, and Probability. *Journal of Experimental Psychology: General, 127*, 269–285.

Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: Sufficiency and necessity priors. *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3110–3115). Austin, TX: Cognitive Science Society.

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

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

Mowshowitz, A., & Dehmer, M. (2012). Entropy and the complexity of graphs revisited. *Entropy, 14*, 559–570.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General, 146*, 1761–1780.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, UK: Cambridge University Press.

Rashevsky, N. (1955). Life, information theory, and topology. *The Bulletin of Mathematical Biophysics, 17*, 229–235.

Rehder, B., & Martin, J. B. (2011). A generative model of causal cycles. *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2944–2949). Austin, TX: Cognitive Science Society.

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

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77*, 427–442.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Vol. Ed.), *The psychology of learning and motivation: Vol. 21*, (pp. 229–261). San Diego, CA: Academic Press.

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

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

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

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater, & M. Oaksford (Eds.). *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 453–484). Oxford, UK: University Press.

Wellen, S., & Danks, D. (2012). Learning causal structure through local prediction-error learning. *Proceedings of the 34th annual conference of the cognitive science society* (pp. 2529–2534). Austin, TX: Cognitive Science Society.

White, P. A. (2002). Causal attribution from covariation information: the evidential evaluation model. *European Journal of Social Psychology, 32*, 667–684.

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

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

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin and Review, 24*, 1488–1500.