Learning Deterministic Causal Networks from Observational Data

Ben Deverett
ben.deverett@mail.mcgill.ca
McGill University

Charles Kemp
ckemp@cmu.edu
Carnegie Mellon University

Abstract

Previous work suggests that humans find it difficult to learn the structure of causal systems given observational data alone. We show that structure learning is successful when the causal systems in question are consistent with people's expectations that causal relationships are deterministic and that each pattern of observations has a single underlying cause. Our data are well explained by a Bayesian model that incorporates a preference for symmetric structures and a preference for structures that make the observed data not only possible but likely.

Keywords: structure learning, causal learning, Bayesian modeling

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 where 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 infer that A sends power to B after observing that A and B are both inactive during some blackouts, that B alone is inactive during others, but that A is never the only inactive station.

A consensus has emerged that causal structure learning is difficult or impossible given observational data alone. For example, Fernbach and Sloman (2009) cite the results of 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 join Mayrhofer and Waldmann (2011) in challenging this consensus. We show that people succeed in a structure learning task when the causal systems under consideration are aligned with intuitive expectations about causality. Previous studies suggest that people expect causal relationships to be deterministic (Schulz & Sommerville, 2006; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), and expect that any pattern of observations tends to be a consequence of a single underlying cause (Lombrozo, 2007). We ask people to reason about systems that are consistent with both expectations, and find that structure learning is reliably achieved under these conditions.

A previous study by White (2006) asked participants to learn the structure of deterministic causal systems from observational data alone. The structures involved were five-node networks where the nodes represented population levels of five different species. 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. Here, however, we demonstrate that both structures considered by White can be reliably learned in the context of the experimental paradigm that we develop.

Given that humans perform well on the structure learning tasks that we consider, it is natural to ask how this performance is achieved. Mayrhofer and Waldmann (2011) propose that learners rely on a "broken link" heuristic and identify the structure that minimizes the number of cases where a cause is present but an effect is absent. They contrast their heuristic-based approach with Bayesian accounts of structure learning that rely on patterns of conditional independence between variables. We propose a Bayesian account that falls in between these two alternatives. Like Mayrhofer and Waldmann, we believe that models which track patterns of conditional independence are often too powerful to capture the inferences made by resource-bounded human learners. Unlike Mayrhofer and Waldmann, we argue that a Bayesian approach is nevertheless useful for explaining why humans succeed in the tasks that we consider. In particular, we show that human inferences are influenced by two factors that are naturally captured by the prior and the likelihood of a Bayesian model—a preference for symmetric structures, and a preference for structures that explain the observed data without needing to invoke coincidences. We demonstrate that incorporating these factors allows a Bayesian model to account for our data better than an approach that relies on the broken-link heuristic alone.

Bayesian Structure Learning

The causal systems that we consider are simple activation networks. Each network can be represented as a graph G which may include cycles. Figure 1a shows one such graph and a data set D generated over the graph. Each row d_i in the data set D represents an observed pattern of activation—for example, the first row represents a case where nodes A, C and D are observed to be active and node B is observed to be inactive. We will assume that each row d_i is generated by activating a randomly chosen node then allowing activation to propagate through the network. For example, Figure 1b shows that if A is the randomly activated node, the final pattern of activation will match the first row of matrix D in Figure 1a.

The activation networks that we consider have three important properties. First, all causal links are generative, and these generative links combine according to an OR function. For example, node C in Figure 1a will be active if node A is

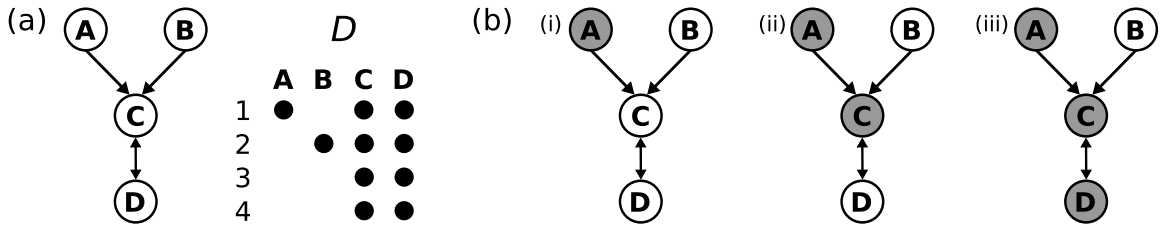

Figure 1: (a) A simple activation network and a data set D generated over the network. Each row of matrix D is an observation which indicates that some of the nodes in the network are active. (b) The first observation in (a) is generated when node A spontaneously activates and activation propagates through the network.

active or if node B is active. Second, all causal links are deterministic. Third, spontaneous activations are rare: at most one node in the network can spontaneously activate at any time, which means that each observed pattern of activation can be traced back to a single root cause. For example, the spontaneous activation of node A is the root cause of the activation pattern in the first row of matrix D in Figure 1a. Our assumptions that causes are rare and have deterministic effects are conceptually related to the work of Lu et al. (2008) on "sparse and strong" priors for causal learning. Note, however, that our notion of rarity is different from their notion of sparsity. Their notion captures the expectation that each node in a causal graph is expected to have at most one strong cause, but ours captures the idea that each pattern of observations d_i is expected to have a single underlying cause. For example, the activation network in Figure 1a is inconsistent with their notion of sparsity, since A and B are both strong causes of C. This network, however, is consistent with our notion of rarity as long as the base rates of A and B are both very low, which means that at most one of these nodes will spontaneously activate at any time.

Because the networks we consider may include cycles, they are different from standard Bayesian networks. If desired, however, our activation networks can be represented as dynamic Bayesian networks where the cycles are unrolled in time (Rehder & Martin, 2011). For our purposes, however, it will be simplest to work with graphs that may include cycles.

Given a data set D generated from an unknown network G, a probability distribution over the possible networks can be computed using Bayes' rule:

$$P(G|D) \propto P(D|G)P(G) = \Big[\prod_i P(d_i|G)\Big]P(G), \qquad (1)$$

where we have assumed that the rows d_i in the matrix D are independently generated over the graph. We will consider two versions of the prior $P(G)$ and two versions of the likelihood term $P(d_i|G)$.

The first version of the likelihood term assumes that observation d_i resulted from the spontaneous activation of a single node in the graph. We sum over all nodes n that may have activated spontaneously:

$$P(d_i|G) = \sum_n P(d_i|G,n)P(n). \qquad (2)$$

$P(d_i|G,n)$ is either 1 or 0 depending on whether d_i is the observation pattern produced by activating node n then allowing activation to propagate through the graph. The prior distribution $P(n)$ is uniform, which captures the assumption that all nodes are equally likely to activate spontaneously. We refer to Equation 2 as the *probabilistic* likelihood.

The second version of the likelihood term depends only on whether observation d_i is consistent with G, and will be called the *logical* likelihood:

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

Observation d_i is consistent with G if d_i can be produced by activating some node in G and allowing activation to propagate through the graph.

The first version of the prior $P(G)$ in Equation 1 corresponds to a uniform distribution over the full space of graphs. The second version captures a preference for graphs that are symmetric. Perceptual research has documented a preference for symmetric stimuli, and this preference can be viewed as an instance of a more general preference for stimuli that display "good form." We hypothesized that a graph shows "good form" if many of its nodes play similar roles. For example, nodes A in B in Figure 1a play similar roles, and exchanging the labels on these nodes leaves the structure of the graph unchanged. The symmetry score of a graph can be formally defined as the number of graph automorphisms, or the number of node permutations that leave the structure of the graph unchanged. For a given number of nodes, the graph with no edges and the fully connected graph will share the highest possible symmetry score, because all possible node permutations leave the structure of these graphs unchanged. We used these symmetry scores to define a prior $P(G) \propto s(G)$, where $s(G)$ is the symmetry score of graph G.

Combining the two likelihoods and the two priors produces a total of four different models. The "logical uniform" (LU) model produces a posterior distribution $P(G|D)$ that assigns equal probability to all graphs G that are consistent with the data. The LU 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. In our setting, a graph is deemed consistent with data D if and only if the graph has a broken link count of zero. When applied to our experimental stimuli, the LU

model therefore makes identical predictions 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.

Like the LU model, the "probabilistic uniform" (PU) model assigns nonzero probability only to graphs that are consistent with the data. The PU model, however, allows for cases where a data set D is consistent with two graphs but better explained by one graph than the other. 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 a causal chain where A sends an arrow to B and B sends an arrow to C. The data, however, are not typical of a chain, since the chain hypothesis requires the assumption that A spontaneously activated 6 times in succession, which seems like a big coincidence. The data are also consistent with a fully connected graph, and now no coincidence must be invoked, since all nodes end up active regardless of which node activates first. As this example suggests, comparing the logical models with the probabilistic models will allow us to evaluate whether people's inferences depend on probabilistic factors like "degree of coincidence" that go beyond consistency with the data.

The "logical symmetry" (LS) and "probabilistic symmetry" (PS) models are directly comparable to the LU and PU models, except that they incorporate a preference for symmetric graphs. Comparing the symmetry models and the uniform models will allow us to evaluate whether people bring *a priori* expectations about the underlying structure to the task of structure learning.

Structure learning experiment

We designed an experiment to explore whether humans are capable of learning the structure of an activation network given observational data alone, and to evaluate the four models just presented.

Participants. 36 members of the CMU community participated in exchange for pay or course credit.

Design. The experiment included 34 blocks, each of which included one or more observations generated over an unobserved network. 32 of the blocks involved networks with three nodes, and the final two blocks involved networks with five nodes. The *characteristic data set* for a network is a set of observations that result from spontaneous activations of each node in the network. Given any network with three nodes, there are 64 possible graphs, but the characteristic data sets for these graphs include only 9 qualitatively different types. Representatives of each type are shown in Figures 2a through 2i. Among the blocks of three-node networks, these nine types were each presented twice, making 18 blocks with three observations each. An additional nine blocks with six observations each were created by including two copies of a characteristic data set per block. Five additional blocks each had two or fewer observations, and are shown in Figures 2j through 2n. These 32 blocks were presented in random order, followed by two final blocks for the five-node networks (Figure 5). These five-node networks are identical to causal structures previously studied by White (2006). The observations within all blocks were shown in random order.

Materials and Procedure. The nodes in each network appeared as rectangles on screen, and active and inactive nodes had different colors. Participants were told that these rectangles were detectors that "detect a rare type of particle called the mu particle." Participants were told that the detectors were connected by directed satellite links, and that an "active detector always activates all detectors that it points to." To reinforce this information, participants were given an example like Figure 1 where they observed a single detector activating and activation subsequently propagating over the network.

Participants then worked through the 34 blocks. Within each block the observations were presented one at a time. After seeing all observations for a given block, participants drew a graph on screen to indicate their best guess about the structure of the underlying network and rated their confidence in their guess on a seven point scale. To minimize memory demands, all observations within a block were retained on screen after being presented, which means that all observations were visible when participants reached the graph-drawing stage. Each previous observation appeared as a panel with detectors, and every edge that participants added during the graph drawing stage was simultaneously added to each of these panels. This design choice was intended to make it as easy as possible for participants to see whether the graph that they had drawn was consistent with all observations for that block.

Results. We focus first on results for the three-node networks. The first nine panels in Figure 2 show the most popular graphs for the nine characteristic data sets, and the remaining panels show results for the blocks with two or fewer observations. In each case the most common response is consistent with the data set, indicating that participants understood the task and were successfully able to discover causal structure given observational data alone. In particular, note that all 36 participants discovered the common effect structure in Figure 2d and the common cause structure in Figure 2f. 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.

Figure 2 also includes predictions of the PS model, and correlations between the model and the data are shown. Results for all four models across the first 32 blocks of the experiment are shown in Figure 3. The first correlation in each panel shows the performance of a model across the entire set of blocks, and the correlation in parentheses shows the average single-block correlation. The PS model performs best overall, suggesting that the probabilistic likelihood and the symmetry prior are both required in order to capture human judgments. A bootstrap analysis indicates that the overall and average single-block correlations achieved by the PS model are reliably higher than the correlations achieved by the PU

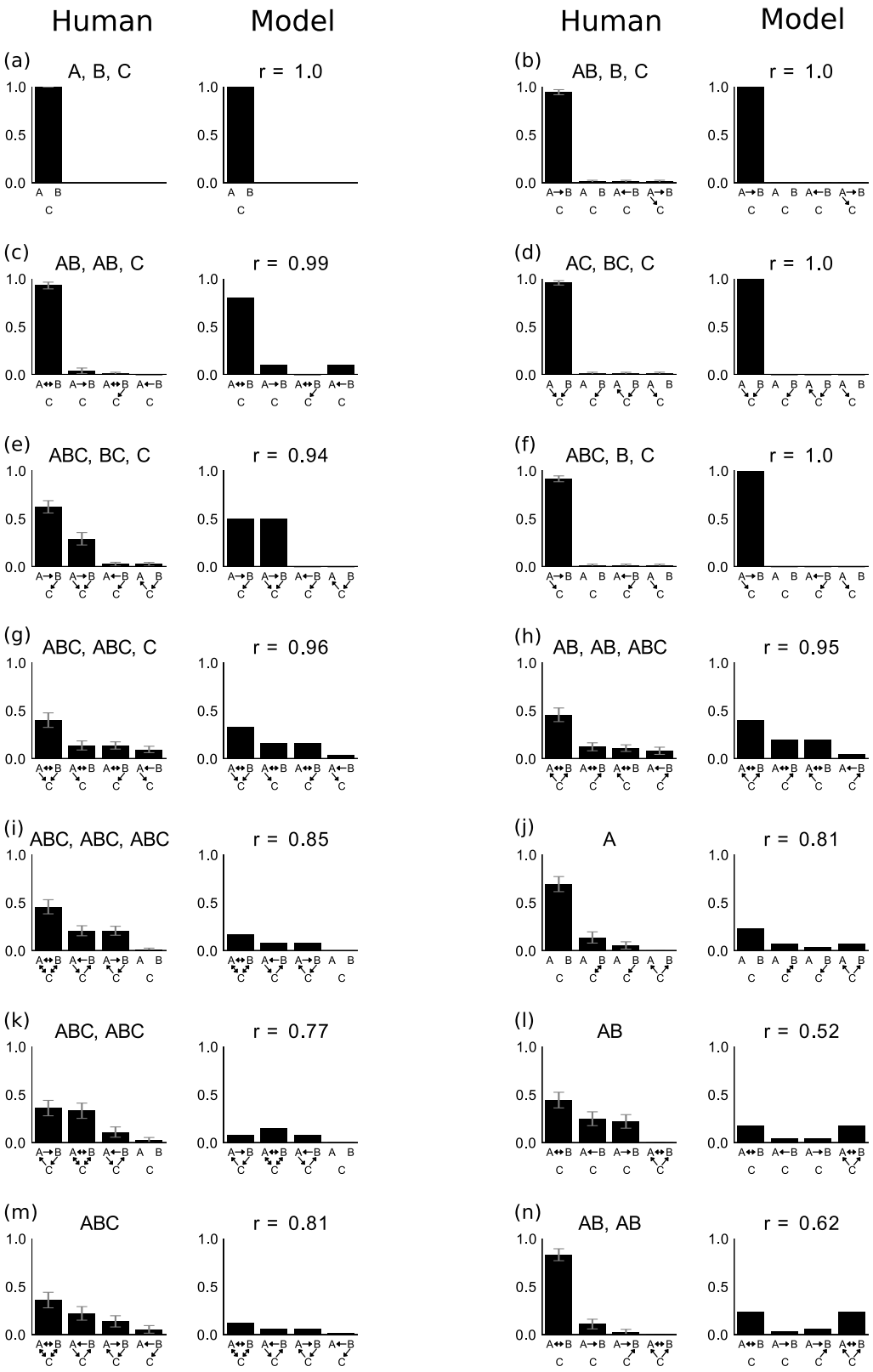

Figure 2: Participant responses and predictions of the PS model for 14 patterns of observations. The observed data are shown above the left plot in each panel, and the correlation between model predictions and human responses is shown above the right plot. The four structures in each plot always include the top two structures chosen by humans and the two most probable structures according to the model.

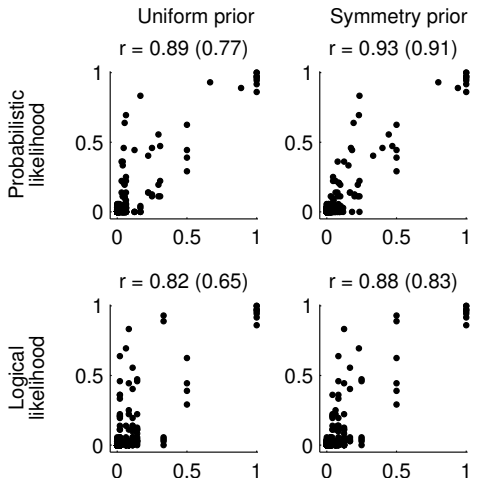

Figure 3: Comparison of the complete set of responses for the first 32 experiment blocks with the predictions of four models. The first correlation in each panel shows correlations based on the complete set of responses, and the correlation in parentheses shows the average correlation across the 32 individual experiment blocks.

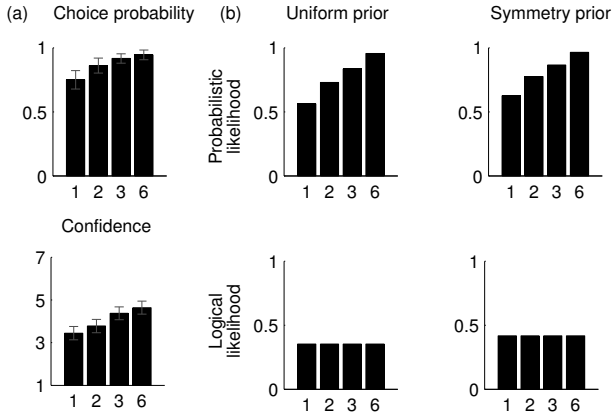

Figure 4: Inferences after observation ABC is presented one, two, three or six times. (a) Proportion of structures chosen by humans that can only generate observation ABC (top); Average confidence ratings (bottom). (b) Probability assigned to structures that can only generate ABC by four models.

and LS models ($p < 0.01$ in all cases).

The main shortcoming of the logical 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 LU model, but the model often assigns the same probability to many other structures. For example, after observing "ABC" three times in succession, the LU model assigns the same probability to all 51 structures that can generate the observation "ABC," including causal chains over these variables. In contrast, the PU model assigns highest probability to the 18 structures that can *only* generate the observation "ABC."

Although the PU model performs better than the LU model, its predictions are still more diffuse than the human responses. As just mentioned, the PU model predicts that 18 different structures are equally likely after observing "ABC" three times, but participants overwhelmingly prefer the top three structures shown in Figure 2i. The symmetry prior allows the PS model to capture this preference: note that the fully connected graph is the most symmetric structure that can only generate "ABC," and the two cycles are the next most symmetric structures that meet this criterion.

To further evaluate the difference between the probabilistic and the logical likelihood, we examined the learning curves that result when the same observation is presented multiple times. The 34 blocks in the experiment include blocks where observation "ABC" is presented once, twice, three times, and six times. Figure 4b shows model predictions for these four blocks, where each bar represents the probability mass assigned to structures that can only generate "ABC." The learning curves for the LU and LS models are flat—these models are sensitive to whether or not a structure is consistent with an observation, but the number of times that the observation appears is irrelevant. In contrast, the PU and PS models become increasingly confident that the underlying structure can only generate the observation "ABC." Figure 4a indicates that participants show a similar learning curve, and become increasingly confident in their responses as the number of observations increases. Bootstrap analyses indicate that the differences between the first and the final bars are statistically significant for both plots in Figure 4a ($p < 0.001$).

Figure 5 shows the most popular graphs chosen for the two five-node blocks. Each set of observations is consistent with only one structure, and participants were reliably able to discover these structures. Figure 6 compares our results to those reported by White, who found that relatively few participants were able to discover these five-node structures. There are at least two reasons why these tasks may have produced different results. First, our particle-detector scenario may be more intuitive than White's task which required inferences about changes in the populations of species over time. Second, we asked participants to reason about the five-node structures following 32 inferences about three-node structures, which means that practice and familiarity with the task may have contributed to their performance. Future studies are needed to isolate the critical differences between these paradigms, but for now we can conclude that there are conditions under which people reliably discover White's five-node structures from observational data alone.

Taken overall our results support two general conclusions. First, humans succeed at structure learning when causes are strong and when each observation has a single root cause. Because our cover story made these conditions quite clear, our data suggest that people reason accurately about deterministic systems where causes are rare, but not that people sponta-

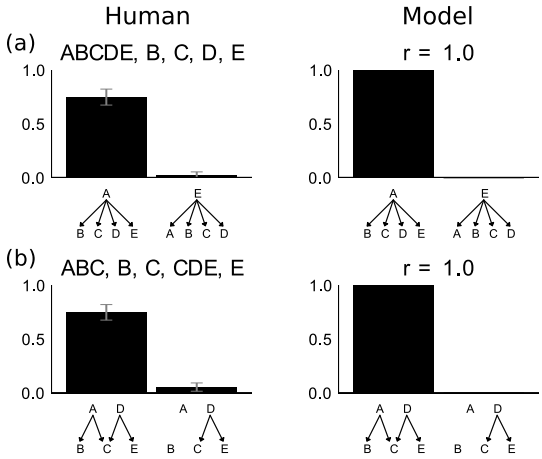

Figure 5: Data sets, human responses and model predictions for the final two experiment blocks. All four models make the same prediction, because in both cases only one structure is consistent with the observations.

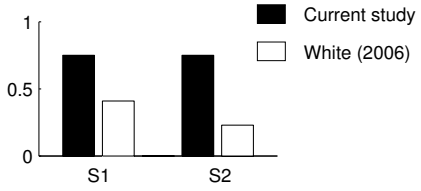

Figure 6: Comparison between our results and results reported by White (2006). The bars show the proportion of participants who successfully learned the five-node structures S1 and S2.

neously bring these assumptions to causal learning problems. Previous studies, however, suggest that both the determinism assumption and the rarity assumption may both apply more generally (Lu et al., 2008; Lombrozo, 2007)

The second general conclusion is that structure learning in our task cannot be adequately characterized as a search for a structure that is consistent with the observed data. At least two additional factors play a role: humans are sensitive to whether or not the observed data are typical of a given structure, and humans have *a priori* preferences for certain kinds of structures including symmetric structures. The PS model illustrates that these factors can be captured by the likelihood and prior of a Bayesian model, and demonstrates the value of the Bayesian approach to structure learning.

Conclusion

Previous studies have found that structure learning from observational data is difficult. In contrast, our data suggest that humans find structure learning relatively easy in settings where causes act deterministically and where each observation has a single root cause. Future studies can consider relaxations of these conditions and explore whether humans still succeed at structure learning when causes are strong but not fully deterministic, and when most but not all observations have a single root cause.

Our data are consistent with the recent work of Mayrhofer and Waldmann (2011), who also report positive results for learning from observational data. Mayrhofer and Waldmann (2011) argue that humans succeed at structure learning by relying on simple heuristics, but we found that their "broken link" heuristic accounted less well for our data than a Bayesian model that incorporates a probabilistic likelihood term and a symmetry-based prior. There may be alternative heuristics that can implement the computations required by our Bayesian model, but we believe that any successful account of our data will need to incorporate an *a priori* preference for symmetric structures, and a preference for structures that make the observed data not only possible but likely.

Acknowledgments. We thank Alan Jern for assistance with the experiment. This work was made possible by a training program in Neural Computation that was organized by the Center for the Neural Basis of Cognition and supported by NIH R90 DA023426.

References

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*(3), 678–93.

Lagnado, D., & Sloman, S. A. (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.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984.

Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: sufficiency and necessity priors. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3110–3115). Austin, TX: Cognitive Science Society.

Rehder, B., & Martin, J. B. (2011). A generative model of causal cycles. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2944–2949). Austin, TX: Cognitive Science Society.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: causal determinism and children's inferences about unobserved causes. *Child Development*, *77*(2), 427–442.

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

White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, *18*(3), 454–480.