Learning Grounded Causal Models

Noah D. Goodman (ndg@mit.edu), **Vikash K. Mansinghka** (vkm@mit.edu), **Joshua B. Tenenbaum** (jbt@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

We address the problem of learning *grounded causal models*: systems of concepts that are connected by causal relations and explicitly grounded in perception. We present a Bayesian framework for learning these models—both a causal Bayesian network structure over variables and the consequential region of each variable in perceptual space—from dynamic perceptual evidence. Using a novel experimental paradigm we show that humans are able to learn grounded causal models, and that the Bayesian model accounts well for human performance.

Keywords: causal variable; causal model; concept grounding; observation; causal learning.

Introduction

Imagine a toddler newly acquainted with the family cat, who soon learns that petting causes purring. *Prima facie* this is a simple inference from observations: the constant conjunction of petting with purring leads to knowledge of the causal relation between them. A number of approaches have made substantial progress in explaining how causal structure can be learned in this way (Cheng, 1997; Waldmann and Martignon, 1998; Gopnik et al., 2004; Sloman, 2005; Griffiths and Tenenbaum, 2005). However, all of these theories share a common limitation: observable causal variables—the states of the world that causal relations relate—must be supplied from the start. Thus, our toddler finds an obstacle to applying any cognitive tools proposed by these theories: what is "petting"? That is, what unifies all of the perceptually different observed instances of petting, and separates them from non-petting—and what makes petting a relevant causal variable? To an infant many aspects of the world might appear as a "blooming buzzing confusion," or at least a largely undifferentiated perceptual space. Yet that perceptual space is soon carved into the separate concepts that become the building blocks of causal understanding—the observable variables. Where do observable variables come from? This question has been raised in recent work on causal learning and cognitive development (e.g. Gopnik et al., 2004), but not addressed in either formal models or behavioral experiments. We begin to do both in the following pages.

It could be that causal variables are not learned at all, but rather we have an innate endowment of variables and must do the best we can with causal relations among them. However, considering the abstract and conceptual nature of many variables, it seems likely that many are learned. If they are learned, perhaps the simplest hypothesis for variable formation is bottom-up perceptual clustering: similar perceptions are grouped together by early perceptual processes and only later become available for causal learning. Indeed, it is well known that the perceptual system can perform sophisticated clustering (e.g. gestalt laws of grouping), so this *bottom-up hypothesis* may be a good description of learning in many cases. For instance, all of the perceptions of "explosions" are quite similar (or anyhow dissimilar to other percepts), so it should be easy to cluster them together into a new variable, which could later be used in causal learning ("gas leaks cause explosions").

However, people often make distinctions which are perceptually quite subtle but causally important. For example, to our toddler petting and pounding the kitty may be perceptually similar—similar enough that they would be clustered together by bottom-up grouping (e.g. simply "touching the kitty"). In actuality, petting and pounding have very different effects (purring versus hissing), and confounding them could lead to problematic results from our toddler. Such examples are the rule rather than the exception: laughing is ever-so-similar to crying, but follows from different causes; pressing the "volume" button on some T.V. remotes is very similar to pressing the "channel" button, but has different effects. These examples suggest, contrary to the bottom-up hypothesis, that causal structure is used to learn observable variables... but how could the causal structure have been learned before the variables? Instead, we must assume in these cases that variables are learned along with causal structure, so that variables and structure may interact and constrain each other.

In this view, three layers of knowledge must be learned simultaneously (illustrated in Fig. 1): First, the number of variables and the possible states of each (e.g. petting and purring, which each have two states—happening or not). Second, the causal structure amongst the variables (e.g. petting causes purring). Third, the *observation function* for each variable, which provides perceptual grounding into the world of the senses: each is a map from perceptions to states of the variable. (Sounds with certain timbre and frequency, for example, might be instances of purring.) Together, the number of variables, their states and observation functions, and the causal structure, constitute a *grounded causal model*.

It is difficult to imagine how complete grounded causal models could spring into being fully formed, Athena-like. Indeed, the great virtue of the bottom-up hypothesis is its comprehensible stages: perceptual clustering processes first form variables, then causal learning mechanisms take over to discover causal structure[1]. We will argue, however, that inference of complete models can be described, at the computational level, in a Bayesian framework. Bayesian induction has been used to describe the learning of causal structure (Griffiths and Tenenbaum, 2005), and it naturally adapts to describe the joint learning of structure and grounding. In the next section we build a model of learning by combining a

[1] We mean in particular bottom-up processes that separate variable formation from causal knowledge, for instance as an informationally encapsulated module.

simple dynamic Bayesian network model of causal structure with a "consequential region" model (Shepard, 1987) of observation functions. Because the number of variables is free to vary, the model can learn as many, or as few, variables as are useful in explaining the perceptual evidence.

After developing this model, we present a simple experimental paradigm to explore grounded causal learning when neither the variables nor the causal structure are known in advance. In this paradigm we describe three conditions, analogous to the petting/pounding example, which are indistinguishable to a bottom-up learner (see Fig. 1). We show that people are able to successfully learn in this situation, and clearly distinguish the three conditions. Further, we find that model predictions correlate well with human responses.

Modeling

In this section we introduce a Bayesian model for learning grounded causal models, together with a specific example situation that forms the basis for our experimental tests. The hypotheses of this model consist of three parts: the number of variables, observation functions for the variables, and a causal structure over the variables. The observation functions provide the relationship between the variables and perception, while the causal structure provides the relationship between the variables themselves. The machinery of Bayesian probability gives a principled method to combine the separate pieces and to draw joint inferences about those pieces, balancing complexity against the ability to explain perceptual data. In the remainder of this section we describe each piece of the model, then assemble the pieces into a likelihood function describing the probability of a sequence of percepts; finally, Bayes' rule is used to invert this dependency, giving the posterior probability of each grounded causal model.

We start with a space of possible perceptual configurations[2] in any given instant, $\mathcal{P}$. An *observable variable* consists of a set of states, S, and an observation function, $f\colon \mathcal{P} \to S$, mapping each point in the perceptual space to a state of the variable. We will focus on the case of binary variables (i.e. $S=\{0,1\}$), in which the observation function is determined by a *consequential region* (Shepard, 1987)—the pre-image of the "active" state (i.e. $f^{-1}(1)$).

Take a simple example world of small dots appearing on a rectangular screen. For a screen with a single dot the perceptual space is given by screen coordinates: $[0,1]^2$. If the number of dots can vary, from none up to a maximum of M, the perceptual space is: $\mathcal{P}=\bigcup_{m=0}^{M}[0,1]^{2m}$. In this world a useful set of observation functions is given by "buttons": rectangular regions that are active when there is a dot inside them. If $\mathbf{r} \subseteq [0,1]^2$ is such a rectangle, and $w=(w_1,\ldots,w_k)\in\mathcal{P}$ is a percept, then the "button" observation function corresponding to $\mathbf{r}$ is $f_{\mathbf{r}}(w) = \bigvee_{m=1}^{k} \delta_{w_m\in\mathbf{r}}$ (that is, 1 if and only if at least one of the k dots is in the rectangle).

Returning to the general situation, say that we have a set of N variables, v_i, with corresponding observation functions f_i. A causal structure C relates the sequence of states $s_{i,t}=f_i(w_t)$ observed from a sequence of percepts $w_t\in\mathcal{P}$. We assume that causal relations hold only between cause variables at one instant and effect variables at the next (that is, we assume that causality is dynamic—causes precede effects—and Markov—states only depend directly on states in the previous instant). Thus the causal structure C is a graph on the variables such that the direct parents of variable v_i, indicated by $\mathrm{par}_C(v_i)$, are its causes. For simplicity we assume a simple nearly-deterministic-or parametrization: a state is active at a given instant, $s_{i,t}=1$, if any parent $v_j\in\mathrm{par}(v_i)$ is active at the previous instant, $s_{j,t-1}=1$, or with some small probability ε (and similarly for off). In addition, any state may be made active by an *intervention*: $s_{i,t}=1$ if $\mathrm{Int}_{i,t}=1$. Putting together this causal structure[3]:

$$P(s_{i,t}=1|s_{t-1},C,\mathrm{Int}) = \begin{cases} \frac{1}{1+\varepsilon} & \text{if } \mathrm{Int}_{i,t}=1, \text{ or} \\ & \exists v_j\in\mathrm{par}_C(v_i) \text{ s.t. } s_{j,t-1}=1, \\ \frac{\varepsilon}{1+\varepsilon} & \text{otherwise.} \end{cases} \tag{1}$$

Next, the observation function f_i of each variable determines a consequential region $f^{-1}(1)$, where the variable is active, and its complement $f^{-1}(0)$, where the variable is inactive. There is a region of perceptual space compatible with any set of states s_t of the variables: $R_{s_t} = \bigcap_i f_i^{-1}(s_{i,t})$. We assume that percepts of a given state are drawn uniformly from this region. Thus the likelihood of a particular percept w_t, conditioned on the observation functions f and states s_t, is:

$$P(w_t|s_t,f) = \frac{\delta_{f(w_t)=s_t}}{|R_{s_t}|}. \tag{2}$$

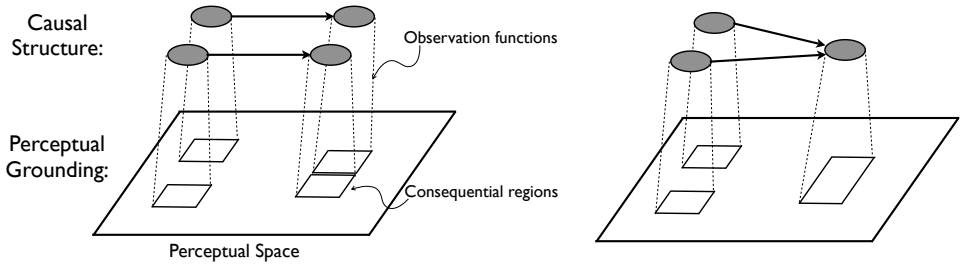
This likelihood leads to a *size principle* (Tenenbaum and Griffiths, 2001): a percept is assigned higher probability when it falls in a smaller region $R_{f(w)}$. Critically, this provides inductive pressure to select variables that minimize the average size of the regions $R_{f(w_t)}$—cutting regions of perceptual space which are frequently visited into finer pieces than those which are rare.

For the example world described above, interventions should be thought of as taps (or mouse-clicks) on the screen, which activate a "button" if they fall within it. We make one minor adjustment to Eq. 2 for this situation: if a variable is made active by an intervention then there must be a dot at the site of the intervention, when this happens $P(w_t|s_t,f)=1$. This has an interesting consequence: a variable which is made active only by interventions (e.g. a variable with no causal parents) will exhibit no size principle effect. We return to this prediction in the next section.

The full likelihood comes from Eqs. 1 and 2, by marginalizing over states, and recalling that each state is independent

[2] We leave open the question of what level of perception this space represents, for instance egocentric or allocentric.

[3] We will simplify subscripts of quantities like state $s_{i,t}$ so that s_t indicates the vector of states at time t, and $\mathbf{s}$ indicates all states at all times.



(a) (b)

Figure 1: Grounded causal models: number of variables, causal structure, and observation functions. Panels (a) and (b) show models that cover the same regions of perceptual space, so could not be distinguished by bottom-up clustering.

of others at the same instant:

$$P(w_t|s_{t-1},C,f,\text{Int}) = \sum_{s_t} P(w_t|s_t,f)\prod_{i=1}^{N} P(s_{i,t}|s_{t-1},C,\text{Int}). \tag{3}$$

However, since Eq. 2 is zero for all but the "observed" state $s^{\text{ob}}_{i,t}=f_i(w_t)$, this simplifies to:

$$P(w_t|s^{\text{ob}}_{t-1},C,f,\text{Int}) = \frac{1}{|R_{s^{\text{ob}}_t}|}\prod_{i=1}^{N} P(s^{\text{ob}}_{i,t}|s^{\text{ob}}_{t-1},C,\text{Int}). \tag{4}$$

For a sequence of perceptions $\mathbf{w}$ with observed states s^{ob}_t (we assume that all initial states $s^{\text{ob}}_{i,\text{-}1}=0$):

$$P(\mathbf{w}|f,C,\text{Int}) = \prod_{t=0}^{T} P(w_t|s^{\text{ob}}_{t-1},C,f,\text{Int}). \tag{5}$$

By Bayes rule, the posterior probability of a grounded causal model f,C is:

$$P(f,C|\mathbf{w},\text{Int}) \propto P(f,C)P(\mathbf{w}|f,C,\text{Int}). \tag{6}$$

For simplicity, and in order to focus on intrinsic aspects of the model, we use a uniform prior on number of variables, causal structure, and observation functions: $P(f,C)\propto 1$. (The uniform prior on number of variables is not proper, but is regularized by the likelihood.) To account for memory limitations on the sequence of percepts, and possible discounting of earlier information, we include a power-law decay in the likelihood term (T is the last time):

$$P(f,C|\mathbf{w},\text{Int}) \propto \prod_{t=0}^{T} P(w_t|s_{t-1},C,f,\text{Int})^{(T-t)^{-\gamma}} \tag{7}$$

Experiment

To investigate human abilities to learn grounded causal models when neither the variables nor the causal structure are known in advance, we adopted the simple perceptual space described above. Participants interacted, by clicking freely, with a blank rectangular window on a computer screen, which they were told was an "alien panel". In response to these interventions a sequence of one or more dots would sometimes bloom and then disappear at various locations on the window. The laws underlying the dots' appearance, unknown to the participants, were similar to the example used earlier: a number of invisible rectangular "buttons" served as variables (the consequential region of each identical to its physical extent on the screen). These buttons were related by a deterministic dynamic causal structure, in which a button was active if any parent was active or if an intervention click was made within it. An active button created a dot within its (invisible) boundary. (As above, a click inside a button always resulted in a dot at the site of the click.)

We wished to create an experimental situation, similar to the petting/pounding example discussed in the introduction, in which two variables had very similar perceptual properties—so would be indistinguishable by bottom-up grouping—but different causal properties. Because the consequential regions of the buttons are rectangular, a set of dots occurring within a single large button occupies the same perceptual space as dots appearing at random in two small buttons that subdivide the large one (see Fig. 1). Thus the three structures illustrated at the left of Fig. 2, which occupy the same regions of perceptual space, should be indistinguishable to a purely bottom-up learner. To put it another way, if one simply clusters the dots appearing within these regions there is no reason to split adjacent rectangles into separate clusters. However, without splitting the adjacent rectangles it is impossible to correctly learn causal structures *a* or *b*. (Of course a clustering algorithm that takes advantage of causal information could distinguish these conditions—the Bayesian model described above can be seen as such an algorithm.)

Thus we had three experimental conditions: one for each structure/condition shown in the left of Fig. 2. With this design we wished to test two main hypotheses: First, people can learn grounded causal models—learning structure and grounding at the same time, and distinguishing conditions impossible for a bottom-up learner. Second, the structure and geometry of the models people infer are consistent with predictions of the Bayesian model.

Method

Participants Participants were 17 members of the MIT community. Two participants failed to understand the instructions, and were excluded from further analyses.

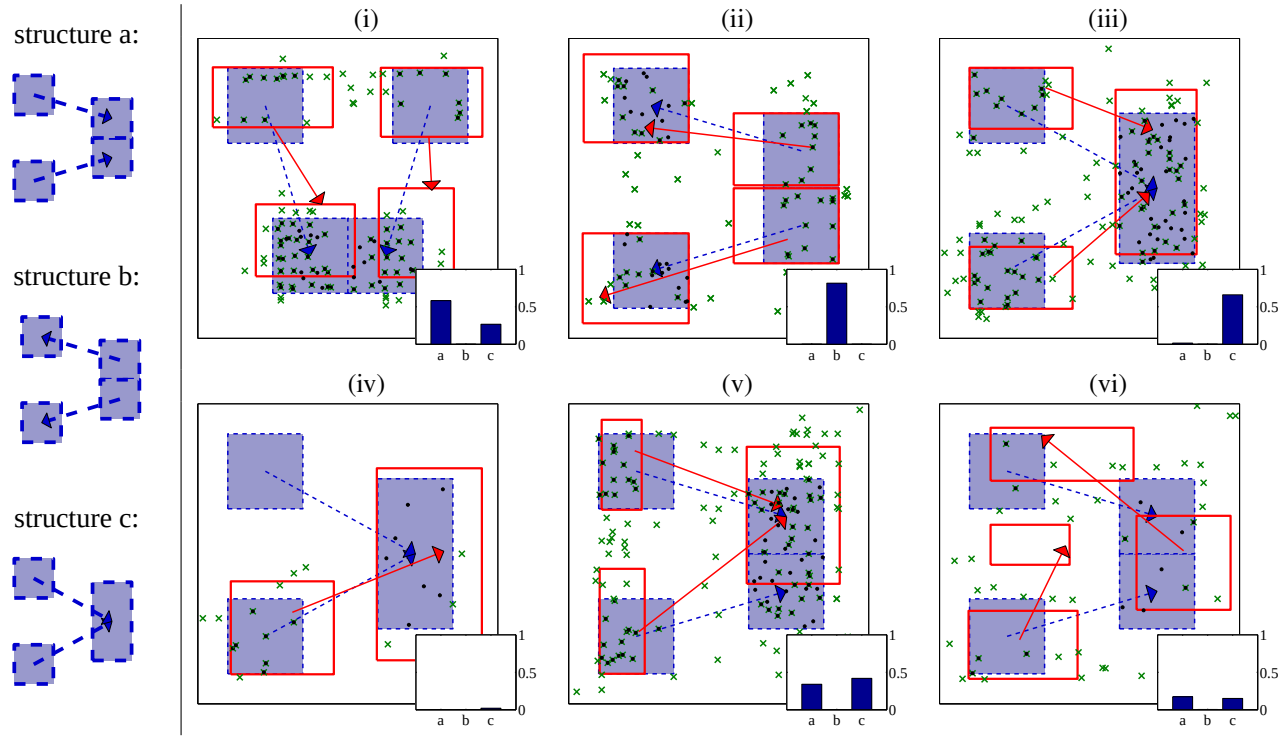


Figure 2: Experimental conditions (structures *a*, *b*, and *c*), and examples of individual participants' observations and responses (i–vi). Participants' interventions are marked by crosses, and resulting dots are marked by dots. Participants' responses are in solid red lines, and the actual structures (not seen by participants) are in dashed blue. Inset in each example is the evidence-specific posterior probability of the model. (i–iii) are correct responses, (iv–vi) are incorrect, as predicted by the model.

Materials and Procedure All interactions, responses, and instructions were presented on a computer screen[4]. Participants first read a brief cover story: "Scientists have discovered alien artifacts that look like blank panels, but respond in interesting ways when clicked upon." Participants were then given a brief familiarization with the "alien panel" screen, and with the drawing tool. The alien panel was blank, except for light grey grid lines, and responded to clicks, as described above, with sequences of blue dots. The drawing tool looked similar, but also had a tool pallet that enabled participants to draw rectangles and arrows. During familiarization participants were told that "scientists have determined that these panels have 'active regions'," and were shown two active regions made visible "by a special process."

There were three "brand new alien panels," one for each of the conditions described above (see Fig. 2, left). Participants interacted with each alien panel for five blocks of approximately 30 clicks each. Each block was followed by a chance to "describe what's going on" by using the drawing tool. Thus each participant made fifteen responses (three conditions, five blocks), each of which was a diagram freely drawn using rectangles and arrows (see Fig. 2, right, for examples of participants' responses). The order of the three conditions was counter-balanced across participants, and the physical orientation of the panels was randomized.

Model approximation and fitting Continuous coordinates were approximated on a fine grid (equal to the pixel width of the screen). Posterior probabilities were evaluated by enumerating over a large subset of hypotheses, including the actual structures and reasonable alternatives that varied in number of variables and causal structure. (Hence the ratios of posterior probability values reported are exact, while the absolute magnitudes are approximate.) The two free parameters of the model, capturing causal transmission noise and "forgetting" rate, were adjusted by hand to ε=0.3, γ=0.5.

Results

We investigate the causal structure and geometry of participants' responses. Before turning to statistical analyses, let us consider a few example responses. In Fig. 2(i-iii) we see the responses of three participants, one in each condition, who correctly identified the underlying causal structure. The rectangular "active regions" are often drawn by participants offset or slightly mis-sized relative to their actual locations. This may be due in part to the difficulty of remembering over the course of the experiment the precise location of interventions and dots. Participants also often separated adjacent variables in conditions A and B, as in Fig. 2(i)—this is suggestive of overregularization due to categorical perception, though could also be due to pragmatic issues with the drawing tool.

[4]An online demonstration of the experiment can be found at: `http://www.mit.edu/~ndg/LGCMdemo.html`

Many "errors" made by participants are, in fact, rational responses to available evidence, and predicted by the Bayesian model. For instance, errors in early blocks are often failures to discover all of the active regions—that is, they are the result of insufficiently broad interventions, see Fig. 2(iv). Occasionally, though all regions have been explored, the evidence generated still favors the wrong hypothesis (according to the model), leading people to a "rational error". For instance, in Fig. 2(v) we see a participant in condition A who has drawn structure *c*—which is also favored by the model (inset). Of course not all errors made by participants are easily explained—Fig. 2(vi).

To enable group analyses, each response was coded as structure *a*, *b*, *c*, or "other" by two coders who were blind to condition and block. (Responses were coded in random order to preclude influence of other responses of the same participant). Coders were instructed to code a response as *a*, *b*, or *c* only if it had the correct number of variables, in approximately the correct locations, and correct causal arrows. There was 96% agreement between the two coders; differences were resolved by discussion. The portion of participants responding with each structure, for each condition and block, are shown in Fig. 3 (black bars). By the final block the modal human response in each condition was correct—that is, the modal response was the actual structure for that condition. Thus participants were able to learn each grounded causal model, given sufficient interaction time. Moreover, participants strongly distinguished the three conditions, providing significantly different response patterns between the conditions ($\chi^2(6)=133.5$, $p<0.001$, aggregating responses across blocks). Since these conditions are indistinguishable to a bottom-up learner, this result makes it quite unlikely that human learning of grounded causal models is purely bottom-up.

We next want to explore whether the structure and geometry of participants' responses are consistent with the Bayesian model. Since participants generated their own interventions, each observed unique evidence; it is thus necessary to apply the model separately to the evidence available to each participant at the time of each response. The inset plots of Fig. 2 show these *evidence-specific* model posterior probabilities. The mean posterior probabilities for each condition and block, averaged across evidence-specific results for individual participants, are shown in Fig. 3 (white bars). Comparing the human response rates (black bars) to the model predictions reveals quite good qualitative agreement. Notice, for instance, that the model correctly predicts that condition A is harder than conditions B or C, and that in condition A a significant minority of responses are structure *c*. The quantitative fit is also good: correlation $r=0.95$, and mean deviation 0.04.

To see whether errors on individual responses were predicted by the model we investigated evidence-specific model posterior probabilities for correct responses (those that matched the actual structure) and incorrect responses. The posterior probability of the actual structure is significantly lower for incorrect trials than correct trials (one-tailed Mann-Whitney U test, $U=2877$, $N=225$, $p<0.0001$). This suggests that the evidence-specific posterior is a good predictor of human errors—and thus that many of these errors are actually rational responses to the available evidence.

Turning to the geometry of the "active regions"—that is, to inferences about properties of the observation functions—we consider the proximity of causes and effects and their relative sizes. In structure *a*, the 'effect' variables were spatially adjacent while the 'cause' variables were separated; vice versa in structure *b*. Participants correctly inferred this aspect of the geometry: in structure *a* responses the 'cause' variables were drawn by participants significantly closer together than the 'effect' variables ($t(56)=5.58$, $p<0.0001$), and in structure *b* responses the 'effects' were significantly closer than the 'causes' ($t(58)=9.38$, $p<0.0001$).

Recall that, because 'cause' variables can only be activated by interventions and an intervention created a dot exactly at the location of the click, only the 'effect' variables should be subject to a size principle under the model. For the 'effect' variables the size principle implies that hypotheses with smaller active regions will have higher probability. Hence, the model predicts the areas for 'cause' variables to be larger, on average, than those for 'effect' variables. This prediction was validated for both structure *a* responses ($t(114)=3.57$, $p<0.001$) and structure *b* responses ($t(118)=2.88$, $p<0.005$).

Discussion and Conclusion

Where do observable variables come from? We have suggested that observable causal variables are learned along with the causal structure that relates them, with causal structure influencing the discovery of variables and variables grounding the causal structure in perception. We developed a Bayesian model to describe the inference of grounded causal models by combining simple Bayesian networks for causal structure with consequential region observation functions. A simple example of this model led to a novel experimental paradigm for the learning of grounded causal models—an aspect of causal learning that has not been previously investigated. In a behavioral test using this paradigm humans were able to correctly learn grounded causal models, and to distinguish conditions that should be indistinguishable to a bottom-up, clusters-then-causes, learner. Further, the model was shown to be a good predictor of the structures and geometry found by human participants, predicting errors on both the group and individual level.

However, our experimental results are only a preliminary test of the model. Further investigation will be needed to evaluate whether learned variables are truly coherent abstract concepts, and to explore the subtle trade-offs between causal structure and observation functions. The model can also be refined in several important ways. First, our simple model of causal structure should be extended to include inhibitory, interactive, and noisy causal relationships. Second, the set

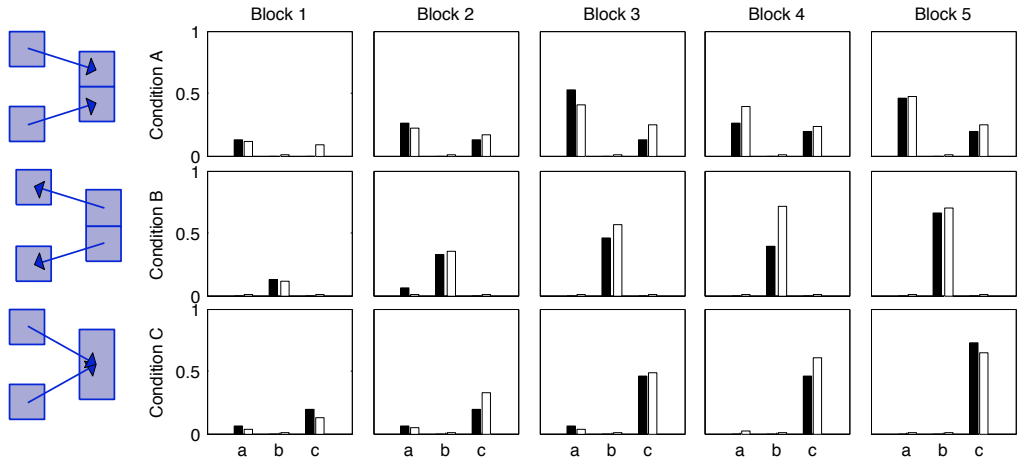


Figure 3: Black bars: the portion of human responses consistent with each structure (*a*, *b*, and *c*; shown at left), for each condition and block. White bars: mean model posterior probabilities. Correlation between the two is r=0.95.

of observation functions we used, while well suited to the experimental situation, is too simple for many real-world settings. It may be possible to adapt existing models of perceptual discrimination, such as fragment-based recognition (Ullman et al., 2001), to provide flexible observation functions. Third, intervention was treated casually; abstract interventions should be grounded in physical actions, and this grounding must also be learned. Beyond simple refinements of this model a more precise and detailed model will eventually be needed: we've only begun to scratch the surface of what could be modeled in this domain—leaving open, for instance, the process and time course of learning, and the generation of exploratory interventions.

We have recently proposed a model for learning causal types (Kemp et al., 2007), an important kind of abstract causal knowledge. (This model was applied it to data from the experiment of Lien and Cheng (2000) in which unknown causal structure interacted with unknown causal types—intriguingly similar to the way that grounding and structure interacted in our experiment.) It is worth noting that the model of Kemp et al. (2007) and the one proposed in this paper can be combined to provide a computational model of causal learning that spans from perceptual grounding to abstract knowledge. Such a combination is especially interesting because it suggests a novel kind of abstract causal knowledge: framework-level prior beliefs about observation functions for new variables. For instance, adults use prior knowledge to quickly infer that a new doorbell-shaped-patch next to a door, even one never before seen, is likely to be a causal variable. Prior knowledge about observation functions is likely at work in many adult inferences of new causal variables.

The ideas of this paper fit in a broader theme in cognitive science and philosophy. In order for concepts to be useful they must have a *semantic grounding*: a manner of connecting an internal representation of the concept to the external world. An influential theory of semantic grounding holds that the meaning of a concept is its observation conditions (ie. states of the world in which it is true). The semantic grounding of concepts has been addressed a number of times in psychology, notably in work on feature formation (Schyns et al., 1998). On the other hand, a great deal of the meaning of a concept could come from its *conceptual role*: the internal relationships it has with other concepts. The theory-theory, especially in its causal models incarnation (Gopnik et al., 2004), has influentially developed this component of concept meaning. By considering *grounded* causal models we have begun to explore a notion of concepts in which meaning comes from both sources: the observation conditions for concepts, and the causal relations between them. By describing how grounded causal models can be learned from perception we have, perhaps, started to unravel how human minds come to have such rich, meaningful, concepts.

References

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, 111(1):3–32.

Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51:285–386.

Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the Twenty-ninth Annual Meeting of the Cognitive Science Society*.

Lien, Y. and Cheng, P. W. (2000). Distinguishing genuine from spurious causes: a coherence hypothesis. *Cognitive Psychology*, 40(2):87–137.

Schyns, P. G., Goldstone, R. L., and Thibaut, J.-P. (1998). The development of features in object concepts (with commentary). *Behavioral and Brain Sciences*, 21:1–54.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237:1317–1323.

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

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–641.

Ullman, S., Sali, E., and Vidal-Naquet, M. (2001). A fragment-based approach to object representation and classification. In *IWVF-4: Proceedings of the 4th International Workshop on Visual Form*.

Waldmann, M. R. and Martignon, L. (1998). A Bayesian network model of causal learning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*.