Learning Causal Laws

Joshua B. Tenenbaum & Sourabh Niyogi
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139
{jbt, niyogi}@mit.edu

Abstract

Attempts to characterize people's causal knowledge of a domain in terms of causal network structures miss a key level of abstraction: the laws that allow people to formulate meaningful causal network hypotheses, and thereby learn and reason about novel causal systems so effectively. We outline a preliminary framework for modeling causal laws in terms of generative grammars for causal networks. We then present an experiment showing that causal grammars can be learned rapidly in a novel domain and used to support one-shot inferences about the unobserved causal properties of new objects. Finally, we give a Bayesian analysis explaining how causal grammars may be induced from the limited data available in our experiments.

Causal Grammars

Recently there has been substantial progress in understanding how people learn causal relations, or causal networks connecting multiple causes and effects. Here we construe *causal network* broadly to include any collection of (domain-specific) causal beliefs that can be represented as a set of nodes and a set of (directed) links between nodes. Nodes may represent objects, properties of or relations between objects, or events. Links may have different causal semantics depending on the semantics of the nodes. For instance, the network N_0 (Figure 1) might represent some aspects of a person's knowledge about several common diseases, their effects (symptoms), and causes (risky behaviors).

Our thesis here is that attempts to characterize people's causal knowledge of a domain primarily in terms of such network structures (e.g., Gopnik & Glymour, 2002; Rehder, in press), while revealing in some important ways, miss a key level of abstraction: the laws that allow people to formulate meaningful causal network hypotheses, and thereby learn and reason about novel causal systems so effectively. For instance, in Figure 1, there appears to be a common domain theory underlying networks N_0, N_1, and N_2, which distinguishes them from N_3, but is not explicitly represented in any of them. We present a framework for representing such abstract causal knowledge, which we call *causal grammar*. The framework is surely incomplete and oversimplified; we view it as merely a first pass at a deep and hard problem. We also describe an experimental study of how people learn and use causal grammars, and briefly sketch a theory of learning based on Bayesian inference.

The networks N_1 and N_2 differ from N_0 in the precise causal links or disease nodes they posit, but they express the same essential regularities: three *classes*

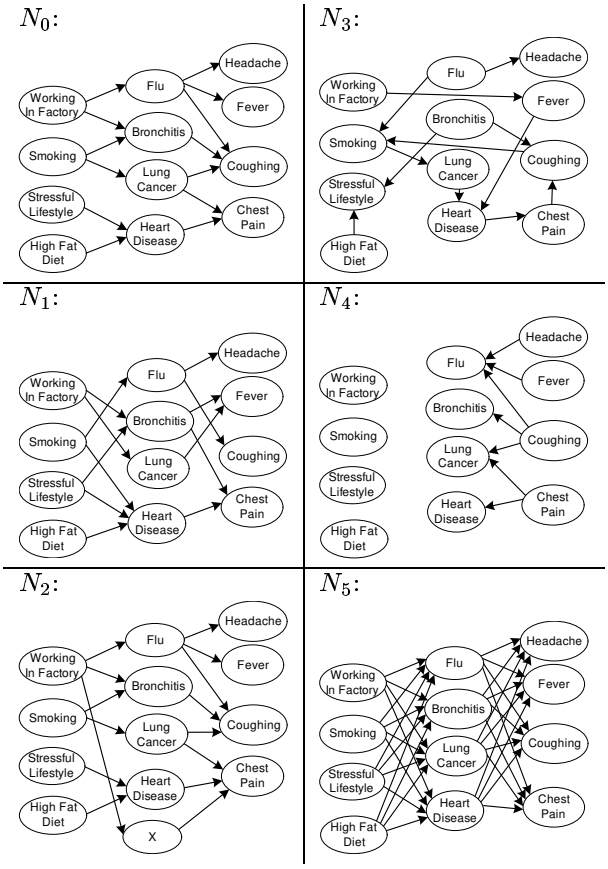


Figure 1: Causal networks in the disease domain.

of nodes (behaviors B, diseases D, and symptoms S), with causal links existing only from behaviors to diseases and diseases to symptoms. A *causal grammar* provides one possible description of this regularity:

G_{BDS}	Node classes:	B, D, S
	Link rules:	$B \rightarrow D$, $D \rightarrow S$

Like a grammar for a language, G_{BDS} specifies abstract classes of entities (nodes, instead of words) and rules about the relations (causal relations, instead of syntactic relations) that may exist between entities of various types. For instance, the rule $D \rightarrow S$ asserts that a causal link may exist from any node in D to any node in S. A causal grammar defines a system for generating causal networks, by first choosing a subset of nodes in each class and then inserting links between those nodes in conformance with the link rules. G_{BDS} generates many causal networks beyond N_0, N_1 and N_2. If node classes may contain infinitely many possi-

ble entities (as seems plausible in the disease domain), the grammar generates an infinite set of networks.

Networks N_3 and N_4 illustrate hypotheses inconsistent with (not generated by) G_{BDS}. N_3 contains the same nodes as N_0, but the links do not respect the division of nodes into three classes. For someone whose beliefs correspond to N_3, not only would we say that he has different beliefs about how particular diseases work, but we would deny that he possesses the same abstract concepts of "disease", "symptom" or "behavior" – at least in the causally relevant sense; he does not know how diseases *in general* work. N_4 appears to respect the same node classes as N_0, but with different link rules: $S \rightarrow D$ instead of $D \rightarrow S$, and no links between behaviors and other nodes.

Causal grammars interact with causal networks through at least three kinds of inferences. Most basic is the causal "parsing" problem. A *parse* of the network N under the grammar G is an assignment of each node in N to some class in G and each link in N to some (consistent) link rule in G. Parsing is an essential (but often implicit) first step before other inferences can occur. For instance, in saying that N_4 is inconsistent with G_{BDS}, we are implicitly parsing the network in a certain way. Parsing ambiguities may arise because the entities corresponding to nodes in a causal network are not generally tagged with class labels. Just as a word may be used in multiple syntactic contexts (e.g., "drink" can be either a noun or a verb), so may a causal node arise in multiple classes (e.g., "overeating" can be either a behavior or a symptom). Even if each node belongs to just one class, when learning about a new domain we may be uncertain about which nodes belong to which classes. This uncertainty is one focus of the experiments below.

Second, causal grammars enable efficient learning and modification of causal networks, by providing crucial constraints on the network hypotheses a learner will consider. To illustrate, consider a learner in the disease domain who has acquired grammar G_{BDS} and currently believes network N_0. She then observes a previously unseen correlation between a known behavior b, e.g., "working in factory", and a known symptom s, e.g., "chest pain". Guided by G_{BDS}, she may infer that a causal chain is likely to go from b to s through some particular but undetermined disease node d. Since no such path exists in N_0, she infers that most likely one of the following new structures is needed: either a link from b to a known cause of s, e.g., "lung cancer", or a new link to s from a known effect of b, e.g., "bronchitis". If no new link to or from an existing disease node can be added without conflicting with other knowledge, G_{BDS} suggests it is likely that a new, previously unobserved disease node x exists, and that x is causally linked to both b and s (as shown in network N_2). Other logically simpler hypotheses, e.g., inserting a single link directly from b to s, or from s to b, are ruled out by the grammar.

Third, causal grammars may themselves be learned or modified based on how well these hypotheses fare – how well they predict the causal networks found in their domain. In its simplest form, the problem of acquiring causal grammars is as follows: we encounter a causal network N generated from some unknown grammar G^* in $\mathcal{G}$, a hypothesis space of candidate grammars, and we must infer the node classes and link rules of G^*. In practice, we often cannot separate this problem from that of inferring the network structure N based on some observed data, guided by our current hypotheses about G^*. The experiment below poses both these challenges simultaneously.

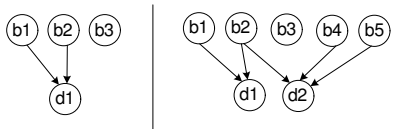
Network N_5 illustrates one ambiguity in inferring causal grammars from network structures. N_5 (unlike N_3) is consistent with G_{BDS}, but differs from N_0, N_1, and N_2 in obeying a stronger regularity: every disease is affected by every behavior and causes every possible symptom. To capture this regularity, we introduce rules for both *necessary* links (denoted $\Rightarrow$) and *possible* links (denoted $\rightarrow$). Then N_5 may be better explained as being generated by the grammar G^*_{BDS}:

G^*_{BDS}	Node classes:	B, D, S
	Link rules:	$B \Rightarrow D$, $D \Rightarrow S$

Later we show how to formalize the preference for one grammar over another in terms of rational statistical inference. Intuitively, G_{BDS} can generate every network that G^*_{BDS} can, plus many more (e.g., N_0, N_1, N_2, etc.). Thus it would be a great coincidence if N_5 had in fact been generated by G_{BDS} and just happened also to be consistent with G^*_{BDS}.

Studies of Acquisition

Our experiments explore the acquisition of a causal grammar in a virtual world loosely inspired by the "blicket detector" studies of Gopnik et al. (in press). In these studies participants are shown a number of blocks, along with a machine called the "blicket detector". The blicket detector "activates" (lights up and makes noise) if and only if a "blicket" is placed on it. Some of the blocks are "blickets", others are not, but their appearance provides no cues. People observe a series of trials in which one or more blocks are placed on the detector and the detector does or does not activate. They are then asked which blocks have the causal power to activate the machine. For instance, they may observe the detector and three blocks, two of which are blickets, as in the left network below:



Nodes correspond to objects, and links correspond to the "activation" relation: $x \rightarrow y$ means that object x activates object y. The task for participants is to infer the link structure of such a network given a series of observations or interventions with the objects. Implicit in this task is a causal grammar that generates the network above:

G_{BD}	Node classes:	B (blocks), D (detectors)
	Link rules:	$B \rightarrow D$

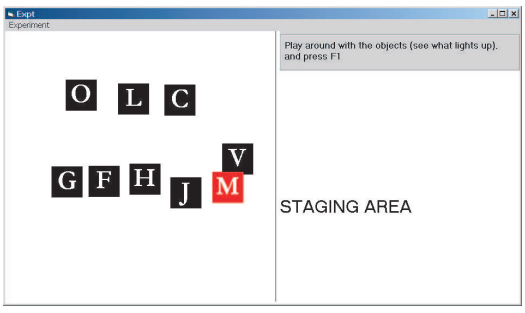
Analogous to rule $D \rightarrow S$ in the disease-symptom domain, G_{BD} specifies that blocks may activate detectors but are not activated by anything, and detectors do not activate anything. One particular element $d1$ of D is called the "blicket detector"; a block b is a "blicket" if there exists a link from b to $d1$. But G_{BD} generalizes beyond the "blicket detector" to generate an infinite number of possible causal networks, relating different kinds of blocks and their detectors. For instance, in the network shown above at right, we have a second detector $d2$; we might call $d2$ the "gazzer detector", and $b2$, $b4$, and $b5$ "gazzers".

Tenenbaum & Griffiths (2003) argue that possession of some abstract theory of detectors, akin to grammar G_{BD}, is necessary to explain people's ability to identify blickets (i.e., acquire the causal network shown above at left) from very few observations. However, no experiments attempt to probe directly whether people in fact have this grammar, or whether they could discover such an abstract system of node classes and link rules if they did not already possess it. Our experiments here begin to probe these questions, using simple grammars such as the following:

G_0	Classes:	A	G_3	Classes:	B, D
	Rules:	$A \rightarrow A$		Rules:	$B \Rightarrow D$
G_1	Classes:	B, D	G_4	Classes:	B, D
	Rules:	$B \rightarrow D$		Rules:	$B \Leftrightarrow D$
G_2	Classes:	B, D			
	Rules:	$B \leftrightarrow D$			

Nodes correspond to objects (square blocks on a computer screen) which are perceptually identical except for a capital-letter label on their face, randomly assigned for reference. A link $x \rightarrow y$ means that when object x touches object y, y will "light up" (turn red). As in the "blicket" domain, activation is an enduring causal power; $x \rightarrow y$ implies that x always lights up y when the two touch. The rule $B \Rightarrow D$ implies that every $b \in B$ activates every $d \in D$. The weaker rule $B \rightarrow D$ implies that any particular b may or may not have the power to activate any particular d. The rules $B \leftrightarrow D$ and $B \Leftrightarrow D$ generate (possible or necessary) bidirectional links: when x touches y, either both objects light up or neither does. The one-class grammar G_0 specifies that any object may activate any other.

In our experiments, participants manipulate a set of objects on-screen whose causal network structure is generated from one of the above grammars. They "pick up" objects with the mouse and touch them together, observing what lights up as in this screen shot:



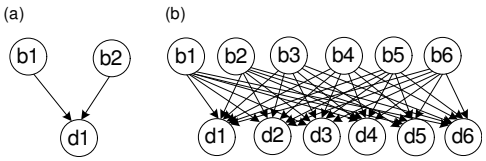


Figure 2: Causal networks generated by grammar G_3 after (a) phase 0 and (b) phase 3 of the experiment.

The experiments comprised multiple phases, during each of which three new objects are added to the on-screen environment. As new objects are added, participants make predictions about how these objects will interact with old objects, allowing us to assess whether they have acquired the correct causal grammar. We also ask them to describe how the objects work, first after seeing just three objects and again after seeing a much larger number (~20).

Crucially, participants are never told about the existence of distinct classes of objects or causal laws defined over those classes. Because there are no static perceptual cues to an object's class membership, the learner faces a "chicken-and-egg" problem: the causal grammar must be discovered (the acquisition problem) simultaneously with inferring which objects belong to which classes (the parsing problem). A real-world analog would be discovering the existence of two classes of magnetic poles ("north" and "south") and the law that poles in different classes attract and those in the same class repel (cf. G_4), simultaneously with inferring which magnets are of which type.

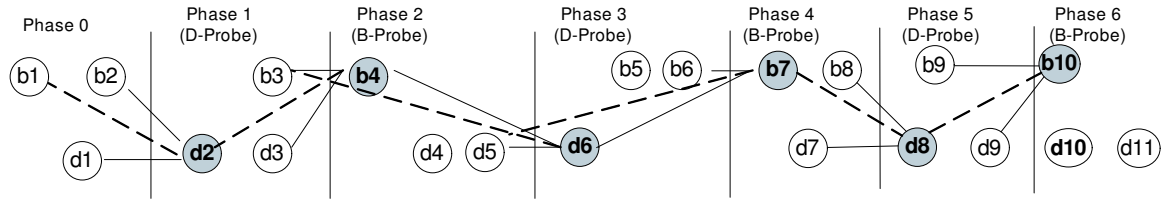
We focus here on one experiment where participants learned the grammar G_3, in which every object in B activates every object in D. As with the law of magnetic poles, this grammar supports very rapid causal inferences: a learner can identify the class of an object just by checking how it interacts with one known B and one known D, and then confidently predict its interactions with every other known object.

Participants. Eleven naive participants were drawn from the general MIT community.

Procedure. The experiment began with phase 0. Three objects appeared in a "staging area", and participants were instructed to "Play around with the objects and see what lights up." The causal network describing the first three objects is shown in Figure 2a. Before proceeding, participants were asked, "Describe how you think the objects work", and typed in a response. Six more phases (1-6) followed, each introducing three additional objects and testing participants' knowledge of the grammar by asking them about the causal properties of a novel object both before and after they had seen it interact with other objects. Each phase had four parts (A through D):

A. Pre-test. Three additional objects were added to the staging area. One new object x served as the "probe". On odd and even numbered phases, the probe was of class D and B respectively. (The other two new objects did not come into play until

(a)

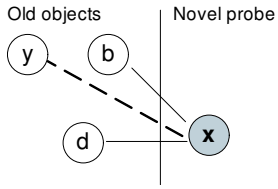


(b)

Question	Prediction (Ideal) Pre	Post	Results Phase 1 Pre	Post	Phase 2 Pre	Post	Phase 3 Pre	Post	Phase 4 Pre	Post	Phase 5 Pre	Post	Phase 6 Pre	Post
Probe $x \in B$														
(1) $x - b$: Will x light up?	5	0			5.1	3.3			5.7	2.0			4.5	0.3
(2) $x - b$: Will b light up?	0	0			3.3	2.7			1.5	1.3			2.4	1.2
(3) $x - d$: Will x light up?	0	0			5.5	3.5			2.9	1.6			2.4	0.5
(4) $x - d$: Will d light up?	5	10			4.7	7.3			4.9	8.3			5.1	8.7
Probe $x \in D$														
(1) $x - b$: Will x light up?	5	10	4.2	7.6			4.4	7.9			5.2	8.8		
(2) $x - b$: Will b light up?	0	0	4.7	4.3			3.6	1.9			1.2	0.6		
(3) $x - d$: Will x light up?	0	0	3.2	6.7			3.1	2.3			2.8	1.2		
(4) $x - d$: Will d light up?	5	0	5.5	4.8			6.5	2.8			4.9	1.9		

Figure 3: (a) Sequence of objects introduced in the experiment. The class of each object is identified by a letter "b" or "d", but participants saw only arbitrary capital-letter labels giving no class information. For each phase, the probe object (x) is shaded; solid lines connect pairs queried in parts A and C of that phase; and a dashed line connects the pair observed to interact in part B. (b) Predictions under the grammar G_3, and participants' mean judgments for pre- and post-interaction questions at each phase, on a 0-10 scale.

part D below.) Participants then received four "pre-interaction" questions, asking about how the novel probe x would interact with two *old* objects, b (in class B) and d (in class D), which they already had experience with. These test interactions correspond to solid lines in the schematic below:



The wording of the 4 questions was as follows:

Consider what will happen when x and b touch.
(1) Will x light up?
(2) Will b light up?
Consider what will happen when x and d touch.
(3) Will x light up?
(4) Will d light up?

(Symbols x, b, and d were replaced by the arbitrary letter labels participants could see on-screen.) Participants responded to each question on a scale from 0 (definitely not) to 10 (definitely). A "successful learner" – a participant who has successfully learned the correct grammar (in this case, G_3) and the class type of each familiar object – should be uncertain about questions (1) and (4) (because the class of x is unknown) but should always answer "definitely no" for questions (2) and (3) (because no b ever lights up and nothing ever lights up when it touches a d).

B. Single interaction. Participants were now instructed to touch the novel probe x to a different familiar object y (connected to x with a dashed line in above schematic) – a single interaction that should allow successful learners to uniquely classify x as a B or D. The target y was always of the opposite type as x, to ensure that in principle x could be uniquely classified from this one interaction.

C. Post-test. Participants answered the same four questions as in the "pre-test" (corresponding to solid lines in above schematic), to assess what they had learned about the causal properties of the probe from the single interaction. Successful learners should now be able to answer all four questions definitively – given causal grammar G_3, they may "parse" the observed interaction of x with the target y and infer the class of x to be B or D; this inference allows them to infer what x can do with other objects. Learners with G_3 can make strong inferences from just a *single observation*; learners without G_3 cannot.

D. Play. Participants were instructed to "Play around with the objects", and proceeded to the next phase at their own pace. By moving objects around, participants could arrange them spatially according to their causal properties. Almost every participant spontaneously discovered some spatial mnemonic (e.g., lining up objects in two rows, one for each class, as in the screen shot above).

Figure 3a shows the class identities of all objects, ordered by phase in which they appeared; the probe in each phase is shaded. After phase 6, participants were again asked to describe how the objects worked.

Results. Figure 3b presents the mean responses to all four questions from each phase of the experiment, both before ("Pre") and after ("Post") the sin-

gle interaction that participants saw between probe x and target y. Also shown are the ideal predictions for a learner who has acquired the correct grammar G_3 and correctly classified all familiar objects. To assess whether, by the end of the experiment, participants had acquired the abstract knowledge necessary to make one-shot causal inferences, we analyzed the differences in their mean pre- and post-interaction predictions from the final phases (5 and 6). In phase 5, given the observation that the probe $d8$ lights up when touching $b7$, successful learners should infer that $d8$ is in class D, and thus increase their confidence that (1) $d8$ will light up when touched to $b8$, and (2) $d7$ will not light up when touched to $d8$. Mean judgments for these two predictions (respectively) changed from 5.2 and 4.9, pre-interaction, to 8.8 and 1.9, post-interaction – both significant differences in the predicted directions ($p = 0.0002, p = .016$ respectively). In phase 6, given the observation that $d8$ lights up when touching the probe $b10$, successful learners should infer that $b10$ is in class B, and thus increase their confidence that (1) $b10$ will not light up when touched to $b9$, and (2) $d9$ will light up when touched to x. Mean judgments for these two predictions (respectively) changed from 4.5 and 5.1, pre-interaction, to 0.3 and 8.7, post-interaction – both highly significant differences in the predicted directions ($p < 0.0001, p = .0007$ respectively).

To assess the course of learning, Figure 4a plots a measure of accuracy versus the number n of objects observed over the six phases. Accuracy was measured in terms of the p-value of the correlation between the ideal predictions (first two columns of Figure 3b) and participants' mean judgments, for both pre- and post-test questions; lower values are better. After the first phase ($n = 3$), the correlation is poor (just above the threshold for significance, $p < 0.05$, dashed line). Accuracy improves rapidly until reaching a plateau below $p < 0.0001$ on the fourth phase ($n = 12$).

There were dramatic differences in participants' verbal descriptions of how the objects work after phase 0 and after phase 6. Descriptions after phase 0 overwhelmingly referenced the particular object labels, e.g., "When F and Z touch, F lights up", with no references to any "group" or "class". In contrast, after phase 6, 9 of 11 participants explicitly used the term "group", "class", or "lighter-uppers" in their descriptions. A typical description was as follows: "XFWIKTAON all light up when touching objects ELHBQRMSDY, but if one box from the first group touches another from the first group then there is no light up. If two boxes from the second group touch [each] other then also there is no light up."

Discussion. Participants clearly learned something like G_3, but given the deterministic flavor of this grammar, it may not be clear why anything more than conventional causal network machinery is needed to describe what people learn or how they learn it. Why not just appeal to standard Bayes net learning algorithms (as in Gopnik et al., in press) to infer a simple deterministic causal network $BD \rightarrow DL$, defined over the variables BD = "an object $b \in B$ touches an object $d \in D$" and DL = "the d object lights up"? This way of thinking locates all of the interesting learning in constructing the variables BD and DL; the trouble is that standard Bayes net learning algorithms take the contents of variables as *given*, and learn only the links between variables. Thus they offer no account of how people discover the existence of the two classes B and D, without which the causal law could not be formulated. More generally, finding satisfying ways to ground the learning of abstract causal knowledge in perceptible relations and attributes (e.g., "touching", "lighting up") is a critical goal for future work.

Bayesian analyses of learning

Participants rapidly acquired a causal grammar from experience, even without any static perceptual cues to objects' class identities, and then used that knowledge to predict the full behavior of a novel object after observing just a single interaction. Most laboratory studies of learning abstract causal knowledge (e.g., Anderson, 1990) have not addressed the problem of simultaneously learning causal laws and the concepts over which those laws are defined, yet much causal knowledge can only be learned in this "chicken-and-egg" fashion. Abstract causal concepts such as "disease" and "symptom" are defined only relationally – diseases are the conditions that cause symptoms – and thus these concepts cannot be understood outside of the causal theory in which they are embedded. However, the theory itself cannot be expressed without reference to these concepts, raising the problem of how either a causal theory or the concepts that comprise it are ever acquired. Learners in our experiment thus face a small-scale version of the problem of conceptual change that takes center stage in "theory theory" accounts of cognitive development (Carey, 1985).

This section sketches a formal analysis of the causal grammar acquisition problem, as posed in our experiments. Learning causal grammars poses an inductive challenge similar to the problem of learning grammars for natural language, in that many grammars are typically consistent with the observed evidence. In our experiment, any network consistent with G_3 is also consistent with G_1, G_0, and numerous grammars with more node classes or link rules. Yet, participants' verbal descriptions and patterns of inference strongly suggest that they acquired the true grammar G_3.

Participants did not show evidence of having acquired the grammar when first tested after encountering just three objects, but appeared to converge on it only after seeing around 12 objects. This progression suggests a kind of statistical inference. Figure 2 shows how the grammar G_3 implies a structural regularity in the causal network – fully connected bipartite structure – that appears highly non-generic after seeing 12 objects (Figure 2b), but not at all remarkable after seeing just three (Figure 2a). We conjecture that people may learn causal grammars in part by detecting these patterns of suspicious coincidence that emerge as more objects are encountered.

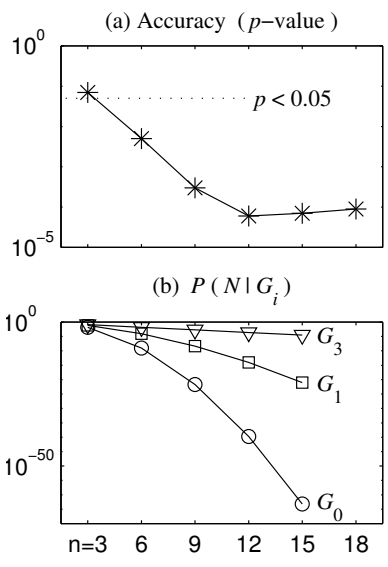


Figure 4: (a) People's accuracy in causal grammar learning increases with the number of objects n encountered. (b) Likelihoods of the observed causal networks under three grammar hypotheses. As n grows, the data become exponentially more likely under the true grammar G_3 than G_0 or G_1.

We can formalize this proposal in a Bayesian framework. The learner has a hypothesis space $\mathcal{G}$ of candidate causal grammars. Each grammar $G_i \in \mathcal{G}$ assigns some likelihood $P(N|G_i)$ to the observed causal network N, and receives some prior probability $P(G_i)$, reflecting how cognitively natural that grammar is. The posterior probability of a hypothesis, $P(G_i|N)$, is then proportional to the product of $P(G_i)$ and $P(N|G_i)$ (Bayes' rule). Many factors could contribute to the prior $P(G_i)$, including both domain-specific knowledge and domain-general preferences for simpler grammars, with fewer nodes classes and link rules. Here we ignore details of the prior and just consider the relative likelihoods $P(N|G_i)$ of the three simplest grammars consistent with the networks in our experiment: G_0, with one node class and one link rule, and G_1 and G_3, each with two node classes and one link rule.

If we make the simplifying assumption that each network consistent with a grammar G_i is equally likely to be generated by it, the likelihoods $P(N|G_i)$ follow the size principle (Tenenbaum & Griffiths, 2001): $P(N|G_i) = 1/\text{size}(G_i)$, for any network N consistent with G_i. The expression size(G_i) denotes the number of networks consistent with grammar G_i. The more expressive the grammar – the larger size(G_i) – the lower the likelihood that G_i would generate any particular network N. In our experiment, these likelihoods are not very informative when only a few objects have been encountered, because the number of causal networks generated by the correct grammar G_3 is not much less than the number generated by alternatives such as G_1 or G_0. For instance, with three nodes, we have size$(G_3) = 7$, size$(G_1) = 13$, and size$(G_0) = 64$. Hence, after seeing only three objects generated from G_3, a Bayesian learner does not yet have strong evidence in favor of the correct grammar.

As the number of objects observed increases, the evidence for G_3 mounts quickly, because the size of the set of networks consistent with G_3 grows much more slowly than for competing hypotheses. On n nodes, G_0 generates all $2^{n(n-1)}$ possible directed networks, and G_1 generates on the order of $2^{n(n-1)/2.5}$ networks (approximately, for $n \leq 10$; more precisely, equal to one less than twice sequence A047864 in Sloane, 2003). In contrast, G_3 generates only the $2^n - 1$ fully connected bipartite directed graphs, which means that G_3 becomes exponentially more likely to have generated the observed data as n increases. At the point where participants' learning asymptotes ($n = 12$), the likelihood $P(N|G_3)$ is more than 10^{10} times greater than $P(N|G_0)$ or $P(N|G_1)$ (Figure 4b).

This Bayesian analysis offers one explanation for how people can infer the correct causal grammar in our experiment, and why they are likely to discover it by the end of our experiment but not after the first phase. We have also extended this approach to studying the learnability of other grammars $(G_0, \ldots, G_4)$ shown above, as well as more complex forms of abstract causal knowledge (Tenenbaum & Niyogi, in preparation). People appear to learn G_4 ($B \Leftrightarrow D$) rather easily, G_1 ($B \rightarrow D$) with some difficulty, and G_2 ($B \leftrightarrow D$) only with great difficulty or not at all. This difficulty ordering is consistent with Bayesian analyses like those above, taking into account the sizes and complexities of alternative grammar hypotheses.

These initial results are promising, but only begin to touch on the questions of how people represent, acquire, and use abstract causal knowledge. We have not investigated or attempted to explain in detail the time course of acquisition. We have also made no attempt to map out systematically the space of cognitively natural causal grammars, or the mechanisms involved in searching this space during learning. These questions loom large for future research.

Acknowledgments. We thank L. Baraff, R. Bryan, and A. Chin for running experiments; M. Bernstein, D. Casasanto, T. Griffiths, R. Kasturirangan, R. Saxe, S. Stromsten, and two referees for helpful comments; and NSF #ECS-9873451 for grant support.

References

J. R. Anderson (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

A. Gopnik & C. Glymour (2002). Causal maps and Bayes nets: a cognitive and computational account of theory-formation. In Carruthers et al. (eds.), *The Cognitive Basis of Science.* Cambridge.

A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, D. Danks (in press). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review.*

B. Rehder (in press). A causal-model theory of conceptual representation and categorization. *JEP: LMC.*

N. J. A. Sloane (2003). The On-Line Encyclopedia of Integer Sequences.

J. B. Tenenbaum & T. L. Griffiths (2001). Generalization, similarity, and Bayesian inference. *BBS* 24:4.

J. B. Tenenbaum & T. L. Griffiths (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems 15.*

S. Carey (1985). *Conceptual Change in Childhood.* Cambridge, MA: MIT Press.