Learning a Theory of Causality

Noah D. Goodman, Tomer D. Ullman, Joshua B. Tenenbaum
{ndg, tomeru, jbt}@mit.edu
MIT, Dept. of Brain and Cognitive Sciences

Abstract

We consider causality as a domain-general intuitive theory and ask whether this intuitive theory can be learned from co-occurrence of events. We begin by phrasing the causal Bayes nets theory of causality, and a range of alternatives, in a logical language for relational theories. This allows us to explore simultaneous inductive learning of an abstract theory of causality and a causal model for each of several causal systems. We find that the correct theory of causality can be learned relatively quickly, often becoming available before specific causal theories have been learned—an effect we term the "blessing of abstraction". We then explore the effect of providing a variety of auxiliary evidence, and find that a collection of simple "input analyzers" can help to bootstrap abstract knowledge. Together these results suggest that the most efficient route to causal knowledge may be to build in not an abstract notion of causality, but a powerful inductive learning mechanism and a variety of perceptual supports. While these results are purely computational, they have implications for cognitive development, which we explore in the conclusion.

Keywords: Causality, hierarchical Bayesian model, innateness.

Introduction

What allows us to extract stable causal relations from the stream of experience? Hume believed that it was the principle of association: constant conjunction of events follow from an underlying association; from this principle, and observed events, one may infer a causal association. Recent research in psychology (?, ?, ?) and philosophy (?, ?) has established the *interventionist* or *causal Bayes nets* (henceforth CBN) account of causality as a description of the principles by which causal reasoning proceeds. These principles include a directed, probabilistic notion of causal dependence, and a privileged role for uncaused manipulation—the *interventions*, which include actions and experimental manipulations. The CBN framework leads to quicker, more reliable learning than weaker assumptions about the nature of causation, and has been successful at predicting human learning (e.g. ?, ?). In recent discussions of the psychological basis of causality it is often assumed that the principles of CBN, or some variant, are an innate resource. In this paper we will argue there is an alternative: that the principles guiding causal understanding in humans can be seen as an *intuitive theory*, learnable from evidence out of more primitive representations. Our argument, which will proceed via an *ideal learner* analysis, can be seen as both an investigation into the psychological basis of causality, and a case-study of abstract learning and the role of innate structure in Bayesian approaches to cognition.

We have previously proposed that intuitive theories—systems of abstract concepts and laws relating them—can be represented in a "language of thought" which includes aspects of probability and logic (?, ?, ?, ?). Because the assumptions of CBN are formalizable via probability and logic, they are potentially expressible in such a language for intuitive theories. This suggests the hypothesis that CBN is not an innate resource, but is itself an *intuitive theory of causality*, learned inductively from evidence and represented in a more basic language of theories.

A theory of causality would have several properties unusual for an intuitive theory. First, it would be domain-general knowledge. Intuitive theories are typically thought of as domain-specific knowledge systems, organizing our reasoning about domains such as physics or psychology, but there is no *a priori* reason to rule out domain-general knowledge. Second, a theory of causality would have to be acquired remarkably early in development. There are intriguing hints that aspects of causal knowledge are present from early infancy, but little evidence that a full notion of cause is innate (?, ?). Yet if a theory of causality is to underly the acquisition of specific causal knowledge it must be available within the first year of life. Could such an abstract theory be learned from evidence so rapidly, even in principle? To investigate this question we turn to hierarchical Bayesian modeling.

The formalism of hierarchical Bayesian modeling makes it possible to express the assumptions relating knowledge at multiple levels of abstraction (?, ?), and Bayesian inference over such a model describes an ideal learner of abstract knowledge (?, ?). Though real learning is undoubtedly resource-constrained, the dynamics of an ideal learner can uncover unexpected properties of what it is possible to learn from a given set of evidence. For instance, it has been reported (e.g. ?, ?) that learning at the abstract level of a hierarchical Bayesian model is often surprisingly fast in relation to learning at the more specific levels. We term this effect the *blessing of abstraction*[1]: abstract learning in an HBM is often achieved before learning in the specific systems it relies upon, and, as a result, a learner who is simultaneously learning abstract and specific knowledge is almost as efficient as a learner with an innate (i.e. fixed) and correct abstract theory. Hierarchical Bayesian models have been used before to study domain-specific abstract causal knowledge (?, ?), and simple relational theories (?, ?). Here we combine these approaches to study knowledge of causality at the most abstract, domain general level.

We will also explore the possibility that learning at the abstract level in an HBM, and the blessing of abstraction, can be substantially aided by providing appropriate low-level features in the input. Our motivation for considering this possibility is a suggestion by ? (?) that part of infants' core knowledge is in the form of *perceptual input analyzers*: mod-

[1] Cf. the "curse of dimensionality".

Law #1:	$\forall x\, \forall y\, A(x) \rightarrow \neg R(y,x)$	Interventions are exogenous.
Law #2:	$\forall x\, A(x) \rightarrow \exists y\, R(x,y)$	Interventions have at most one child.
Law #3:	$\forall x\, F_1(x) \rightarrow A(x)$	Fetaure 1 is diagnostic for interventions.
Law #4:	$\forall x\, F_2(x) \rightarrow A(x)$	Fetaure 2 is diagnostic for interventions.
Law #5:	$\forall x\, \forall y\, R(x,y) \vee R(y,x) \vee x{=}y$	Dependence graph is fully connected.
Law #6:	$\forall x\, \forall y\, \neg R(x,y)$	Dependence graph is unconnected.
Law #7:	$\forall x\, \exists y\, R(x,y)$	Variables have at most one child.
Law #8:	$\forall x\, \exists y\, R(y,x)$	Variables have at most one parent.
Law #9:	$\forall x\, \forall y\, \forall z\, R(x,y) \wedge R(y,z) \rightarrow R(x,z)$	Dependence graph is transitive.
Law #10:	$\forall x\, \forall y\, A(x) \rightarrow \neg R(x,y)$	Interventions have no children.
Law #11:	$\forall x\, \exists y\, \neg A(y) \wedge R(y,x)$	Variables have at most one non-intervention parent.

Figure 1: Eleven laws that can be expressed in the language for theories.

ules that perform simple transformations of raw perceptual input, making it suitable for conceptual cognition. We hypothesize that these perceptual input analyzers do not provide abstract conceptual knowledge directly, but instead serve to make latent abstract concepts more salient and thus more learnable. For instance, the feeling of self-efficacy, advocated by Maine de Biran as a foundation of causality (see discussion in ?, ?), could be an analyzer which highlights events resulting from one's own actions, making the latent concept of intervention more salient. Altogether this suggests a novel take on nativism—a "minimal nativism"—in which strong, but domain-general, inference and representational resources are aided by weaker, domain-specific perceptual input analyzers.

In the following sections we first formalize aspects of CBN within a logical language for intuitive theories. We then study the ideal learner of causal knowledge, investigating the speed of learning at different levels of abstraction, and the effect of perceptual input analyzers on learning speed.

Theories of causality

Causality governs the relationship between events. Formalizing this, the world consists of a collection of causal systems, in each causal system there is a set of observable *causal variables*. Causal systems are observed on a set of *trials*—on each trial, each causal variable has a value. (We will call an observation of a causal variable on a particular trial an event.)

The causal Bayes nets theory of causation (?, ?) describes the structure of dependence between events, isolating a special role for a set of interventions. CBN can be seen as a collection of assumptions about causal dependence: (CBN1) Dependence is directed, acyclic, and can be quantified as conditional probability. (CBN2) There is independence / indirect dependence. (CBN3) There is a preferred set of variables, the "interventions", which are outside the system—they depend on nothing. (CBN4) Interventions influence only one variable. (CBN5) The intervention set is known for each causal system. In addition, assumptions are often made about the functional form of dependence (for instance, that interventions are "arrow breaking"). For simplicity we will address only the aspects of this theory that determine the structure of the dependency relation and will assume (CBN1).

A language for theories of causal dependence

We wish to specify a hypothesis space of alternative theories of the dependency relation, R. This space should contain CBN and a wide set of alternative theories, and should build these theories by combining simple primitive units. ? (?) proposed a very flexible language for expressing relational theories, which is a small extension of first-order logic, and used this language to predict the inductive generalization of human learners in a novel domain. We propose that a version of this language can be used to capture domain-general knowledge, including (aspects of) a theory of causality.

The language we use contains logical operators: quantifiers over causal variables—"for all" ($\forall$), "there exists" ($\exists$), and "there exists at most one" ($\nexists$)—and logical connectives—not ($\neg$), and ($\wedge$), or ($\vee$), if ($\leftarrow$). In addition to the logical operators, and the causal dependence relation $R(\cdot,\cdot)$, the language contains invented predicates and observed predicates. Invented predicates are not observable, or pre-defined, but have a conceptual role in the theory. We restrict in this paper to at most one invented predicate, $A(\cdot)$; this predicate need not *a priori* relate to causality in an interesting way, but it will play the role of defining intervention in the correct theory. Finally, the two predicates, $F_i(\cdot)$, are observable features of variables. These can be thought of as perceptual input analyzers extracting some feature of events[2], which may or may not be useful in a theory of causality.

Fig. **??** gives examples of laws that can be expressed in this language. These laws include those needed to capture the CBN theory of causality, as well as a variety of plausible variants describing alternative restrictions on dependency. Within this set of laws (CBN3) corresponds to Law #1; (CBN4) corresponds to Law #2; (CBN5) follows from Laws #3 and/or #4 when the features can be used to identify interventions; (CBN2) is the lack of Laws #5 or #9.

[2]It is most realistic to think of input analyzers operating at the level of specific events; we idealize them as features of causal variables (i.e. types of events).

A hierarchical Bayesian model

To ground this language for theories into observed events in a set of causal systems, we construct a hierarchical Bayesian model with theories of causality at the most abstract level and events at the most specific level (Fig. **??**). We first describe the generative process of this model, then we describe the ideal learner by inverting this process using Bayes' rule.

Generating a theory A causal theory—represented in the theory language described in the previous section—is drawn from the prior distribution over theories, $P(T)$. We take $P(T)$ to be uniform over theories (of size less than some maximum). While a representation-length prior (see ?, ?) would naturally capture a bias for simpler theories, we choose a uniform prior in order to focus on the dynamics of learning driven entirely by the hierarchical setup.

Generating causal models Next a causal model is generated for each causal system s. A causal model is an instantiation of each predicate in the theory—R_s and, if it is used, A_s. Following (?, ?), we will assume that the distribution on causal models, $P(A_s, R_s|T)$, is uniform over those consistent with T—that is, the instantiations of R_s and A_s that satisfy the logical laws of T.

Generating events Each causal model in turn generates observed events (a value for each variable) for a set of trials. The probability of generating a series of trials $D = \{d_t\}$ from a system with causal relation R is given by:

$$P(D|R) = \int \prod_t P(d_t|R,\Theta)P(\Theta|\alpha)d\Theta \tag{1}$$

Where the *conditional probability tables*, Θ, list the probability of each event given each set of values for its parents in R. We make the weak assumption that each entry of Θ is drawn independently from a symmetric-beta distribution with hyperparameter α. The integral in Eq. **??** is a product of standard beta-binomial forms, which can be integrated analytically.

Theory induction

The ideal Bayesian learner infers a posterior belief distribution over theories from a set of observed trials across a range of causal systems. The posterior probability of a theory, T, given data, $\mathbf{D} = \{D_s\}$ is given by:

$$P(T|\mathbf{D}) \propto P(\mathbf{D}|T)P(T) \tag{2}$$

Where the likelihood is given by:

$$\begin{aligned} P(\mathbf{D}|T) &= \prod_s P(D_s|T) \\ &= \prod_s \sum_{A,R} P(D_s|A,R)P(A,R|T) \\ &= \prod_s \sum_{A,R} P(D_s|R)P(A,R|T) \end{aligned} \tag{3}$$

Figure 2: The hierarchical Bayesian model, and examples of the information at each level. The causal dependence relation $R(\cdot,\cdot)$ is shown as directed edges between variables (circles), the latent predicate $A(\cdot)$ is shown as shading of the variables. Binary events for each system, trial, and variable are shown as contingency tables.

System marginals The effect of an abstract theory on learning in a specific system, s, may be described by the posterior belief distribution over R_s. If we fix a theory, T, and use this to provide the prior over R_s, the posterior is given by:

$$P(R_s|T,D_s) \propto P(D_s|R_s)P(R_s|T) \tag{4}$$

If the theory is not fixed, but is learned simultaneously with the causal systems, we may still want to capture what has been learned about one specific system within the hierarchical setup. This is given by the posterior marginal of R_s:

$$\begin{aligned} P(R_s|\mathbf{D}) &= \sum_T P(R_s|T,\mathbf{D})P(T|\mathbf{D}) \\ &= \sum_T P(R_s|T,D_s)P(T|\mathbf{D}) \end{aligned} \tag{5}$$

Ideal learner simulations

To investigate the dynamics of learning in the theory induction framework outlined above, we performed a series of simulation studies.

The probability landscape of this model is complex, making it difficult to accurately characterize learning at all levels of abstraction. To ensure correct results, we chose to implement the learning model by explicit enumeration over theories and causal structures. To make this enumeration possible we restricted to theories which can be formed as a conjunction of at most five of the laws shown in Fig. **??**, and to systems of only four variables. (Counting only theories with a satisfying causal model, there are 691 theories in the set we considered. There are 543 possible causal structures R, and 16 possible intervention sets A.)

For each run of the model we generated evidence for the learner by first choosing one variable in each system to be

(a)

(b)

(c)

Figure 3: (a) Rank of the correct theory, mean and 10th/90th percentiles across 100 model runs. (b) Rank of the correct causal structure (mean over systems and runs), given no theory, fixed correct theory, and simultaneously learned theory. Learning abstract knowledge always helps relative to not having a theory, and is quickly as useful as an innate, correct theory. (c) The probability of correct learning: the fraction of systems in which the correct structure has been learned (is at rank 1), and the fraction of runs in which the correct theory has been learned. (In each run there were 50 systems, and one feature perfectly diagnostic of interventions. Hyperparameter α=0.5.)

an intervention, then generating a causal model for each system (consistent with the correct, CBN, theory of causality) and data for each trial according to the generative process described above. We initially fixed the number of systems to 50, and included one feature which correlates perfectly with intervention and another which is uncorrelated with intervention; we consider the effect of varying these conditions below.

We explore the dynamics of learning by varying the amount of evidence given to the learner, as measured by the total number of samples (i.e. trials) across all systems, with each system given the same number of samples. The ideal Bayesian learner is able to learn the correct theory, given sufficient evidence (Fig. **??**a). This, by itself, is unsurprising—indeed, Bayesian induction is guaranteed to converge to the correct hypothesis in the limit of an infinite amount of evidence. It is more interesting to see that learning the correct theory appears relatively quick in this model (being achieved with fewer than 30 samples per system in most runs).

The blessing of abstraction

Abstract knowledge acts as an inductive bias, speeding the learning of specific causal structure. Fig. **??**b shows the mean rank of the correct causal structure across systems with no abstract theory (i.e. a uniform prior over causal relations), with innate (i.e. fixed) correct theory, and with learned theory (i.e. with the theory learned simultaneously with specific causal models). We see, as expected, that the correct abstract theory results in quicker learning of causal structure than having no theory. Comparing the learned-theory curve to the no-theory curve, we see that abstract knowledge helps at all stages of learning, despite having to learn it. Comparing the learned-theory curve with the innate-theory curve shows that by around 60 samples per system the theory learner has matched the performance of a learner endowed with an innate, correct theory. Thus, the abstract layer of knowledge can serve a role as inductive bias even when the abstract knowledge itself must be learned—learning a theory of causality is as good (from the perspective of causal model learning) as having an innate theory of causality.

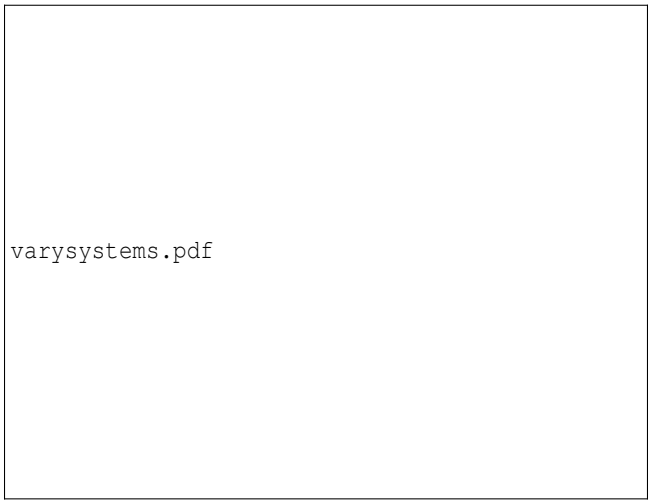

Figure 4: The posterior log-probability of the correct theory as a function of total number of samples across systems, for different numbers of systems. Each curve starts at 2 samples per system. Learning is best when evidence is gathered from many systems, even when only a few samples are taken in each system.

How can abstract knowledge appropriately bias specific learning, when it must be learned itself? Comparing Fig. **??**a to Fig. **??**b suggests that the correct theory is learned before most of the correct causal structures. In Fig. **??**c we have investigated this by plotting the probability of learning (defined as the correct hypothesis being most probable), at the levels of both systems and theories. We see that learning at the abstract theory level is much faster than at the system level. Further, the time to correct learning at the system level is almost identical for innate-theory and learned-theory, which are both faster than no-theory. This illustrates the fact that ab-

Figure 5: Learning curves for the eleven laws of Fig. **??**.

stract learning is not bottom-up, waiting on specific learning; instead, learning is being carried out at all levels simultaneously, and here abstract knowledge is often learned before specific knowledge. Note that this effect is not due to the relative size of hypothesis spaces (we consider 691 theories and 542 specific causal structures), nor a "helpful" choice of prior (we use a maximum entropy—uniform—prior on theories, and for the no-theory case a similar prior on specific systems). Rather, this effect is driven by the ability of the higher level of the model to learn from a wide range of evidence drawn from multiple systems.

To confirm that breadth of evidence is important, we consider the effect of distributing the same amount of evidence among a different number of systems—is it better to spend effort becoming an expert in a few systems, or to be a dilettante, learning only a small amount about many systems? Fig. **??** shows the result of varying the number of systems, while matching the total number of samples (resulting in differing numbers of samples per system). Learning is fastest when evidence is drawn from a broad array of causal systems, even when only a few samples are observed in each system. Indeed, at one extreme learning is very slow when only five systems are available. At the other extreme, learning from 500 systems is quick overall, and "catches up" to other conditions after only three samples per system.

Turning to the dynamics of learning for individual laws, Fig. **??** shows the marginal probability of each of the eleven laws in Fig. **??**. Law #3, relating interventions to the observed predicate F_1, is learned first, but is closely followed by Law #1, which defines the main role of interventions in CBN. Slightly later, Law #2—specifying that interventions effect only one variable—is learned. All other laws slowly drop off as the correct theory becomes entrenched. The gradual learning curves of Fig. **??**, which are averaged over 100 runs of the model, belie the fact that learning of the laws was actually quite abrupt in most runs. Though the exact timing of these learning events was distributed widely between runs, the order of acquisition of the laws was quite consistent: in two-thirds of runs Laws #1 and #3 were learned almost simultaneously, followed later by Law #2. (To be precise, in 92% of runs Law #2 was learned last, as measured by number of samples required to cross probability 0.75; of these runs, Law #1 led Law #3 on 59% of runs, but the two laws were learned within one step of each other on 74% of runs.) This observation may be significant given that cognitive development is characterized by wide variation in timing of acquisition, but remarkable consistency in order of acquisition.

A minimal nativism

Thus far we have assumed that there is an observed feature which can be used to tell when a variable is an intervention. We can imagine that this feature provides information extracted from perception of the observed events—that is, it results from an *input analyzer* (?, ?): an innate mechanism that performs simple transformations of perceptual evidence. A number of relatively simple input analyzers could provide features useful for identifying interventions. For instance, the feeling of self-efficacy discussed by Maine de Biran, or, more broadly, an innate agency-detector able to identify the actions of intentional agents (see ?, ?). Critically, none of these simple input analyzers is likely to identify all interventions (or even most), and they are likely to be mixed together with features quite un-useful for causal learning.

Figure 6: The marginal probability of the correct theory of intervention (i.e. Laws #1 and #2) given different sets of "input analyzers": each condition has two features which are diagnostic of intervention variables to the extent indicated (e.g. "F1 50%" indicates that the first feature covers half of interventions). In the 50%/25% case the two features overlap, otherwise they are disjoint. Learning is difficult when no diagnostic features are present, but quite rapid under all other conditions.

We simulated learning under several different "input analyzer" conditions varying in: the number of useful features (the remaining feature(s) were distractors), what portion of intervention variables could be identified from the useful features, and the overlap between features. In Fig. **??** we have

plotted the marginal probability of the "intervention" portion of the correct theory—Laws #1 and #2, which govern the role of interventions in determining causal dependency, independent of the identification of interventions. We see that learning is extremely slow when no features are available to help identify interventions. In contrast, learning is about equally quick in all other conditions, depending slightly on the coverage of features (the portion of interventions they identify) but not on how this coverage is achieved (via one or multiple features). Thus, even a patchwork collection of partial input analyzers, which pick out only a portion of intervention variables, is sufficient to bootstrap abstract causal knowledge; learning can be relied on to pick the useful features from the distractors and to sort out the underlying truth that each partially represents.

Discussion and conclusion

We have studied an ideal Bayesian learner acquiring aspects of a domain-general intuitive theory of causality. This theory and a wide set of alternatives were represented in a "language of thought" for relational theories, based upon first-order logic. We found that the correct theory of causality can be learned from little evidence, often becoming available before specific causal models have been learned. This enabled the learned abstract knowledge to act as an inductive bias on specific causal models nearly as efficiently as an innately specified theory. However, this "blessing of abstraction" itself relied on a set of observed event features that served to make the latent concept of intervention more salient.

The abstractness of a theory of causality proved not to hinder learning, given a rich language of thought and a powerful inductive learning mechanism. We found that abstract learning was fastest when evidence was drawn from a wide variety of causal systems, even if only a small number of observations was available for each system. Because a domain-general theory is able to draw evidence from the widest set of experiences, this suggests that domain-general intuitive theories may, in some cases, be easier to learn than their domain-specific counterparts. In future work we plan to investigate further the effects of distribution and variety of evidence.

Though we have argued that causality may be learnable, our results should not be taken to support an entirely empiricist viewpoint. We endow our learner with a rich language for expressing theories and a strong inductive learning mechanism. These are both significant innate structures, though ones that may be required for many learning tasks. In addition, we have shown that the domain-general mechanisms for learning and representation are greatly aided by a collection of domain-specific "perceptual input analyzers". It may be ontogenetically cheap to build innate structures that make some intervention events salient, but quite expensive to build an innate abstract theory (or a comprehensive analyzer). Since a powerful learning mechanism is present in human cognition, the most efficient route to abstract knowledge would then be by bootstrapping from these simple, non-conceptual mechanisms. Thus we are suggesting a kind of minimal nativism: strong domain-general inference and representational resources, aided by weak domain-specific perceptual input analyzers.

Our results are purely computational, at the level of ideal learning, but they provide a viewpoint that we believe will be useful for empirical research in cognitive development. When young infants behave as if they have a piece of abstract knowledge, it is tempting to conclude that this knowledge is innate. This tendency may misguide—we have shown that abstract knowledge of causality, at least, can be learned so quickly that it might seem to be innate. On the other side, where innate structure *is* required to explain complex cognition, it is often assumed to be abstract conceptual knowledge (?, ?). This should also be approached with care—simpler innate structures, without conceptual content, may be sufficient when paired with a powerful learning mechanism. Finally, the most obviously acquired systems of conceptual knowledge are coherent explanations of a single domain, yet it may often be the broader domain-general intuitive theories which are acquired earliest and are most fundamental.