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

The very early appearance of abstract knowledge is often taken as evidence for innateness. We explore the relative learning speeds of abstract and specific knowledge within a Bayesian framework, and the role for innate structure. We focus on knowledge about causality, seen as a domain-general intuitive theory, and ask whether this knowledge 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 "perceptual 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.

Pre-print June 2010—to appear in Psych. Review.

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 (Hume, 1748). Recent psychological research (Cheng, 1997; Waldmann & Martignon, 1998; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Gopnik et al., 2004; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Griffiths & Tenenbaum, 2005, 2009) has described mathematical models of how children and adults learn domain-specific causal relations by applying abstract knowledge—knowledge that describes how causal relations give rise to patterns of experience. None of this work, however, addresses a more fundamental question: what are the origins of our general understanding of causality—of the abstract principles by which observation and intervention can be used to infer causal structure? Most previous work has assumed explicitly or implicitly that the human sense of causality is innate, either because it is a necessary building block in models of cognitive development or because it was not clear how such abstract knowledge could be learned. Those who have proposed that our concept of cause is constructed from experience (e.g. Carey, 2009) have not attempted to give a formal learning model. In this paper we will argue 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.

A long tradition in psychology and philosophy has investigated the principles of causal understanding, largely converging on the *interventionist* or *causal Bayes nets* account of causality (Pearl, 2000; Woodward, 2003) as a description of the principles by which causal reasoning proceeds. The principles embodied by the causal Bayes network framework include a directed, probabilistic notion of causal dependence, and a privileged role for uncaused manipulation—the *interventions*, which include actions and experimental manipulations. The causal Bayes network framework leads to quicker, more reliable learning than weaker assumptions about the nature of causation, and has been successful at predicting human learning data (Cheng, 1997; Waldmann & Martignon, 1998; Steyvers et al., 2003; Gopnik et al., 2004; Lu et al., 2008; Griffiths & Tenenbaum, 2005). 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 (Kemp, Goodman, & Tenenbaum, 2008; Tenenbaum, Griffiths, & Niyogi, 2007; Goodman, Tenenbaum, Griffiths,

Acknowledgements: We would like to thank Jim Woodward, Mike Oaksford, and an anonymous reviewer for helpful comments and discussion. This work was supported by: the J.S. McDonnell Foundation Causal Learning Collaborative Initiative, Office of Naval Research grant N00014-09-0124, Air Force Office of Scientific Research grant FA9550-07-1-0075, and Army Research Office grant W911NF-08-1-0242.

& Feldman, 2007). Because the assumptions of causal Bayes networks are formalizable via probability and logic, they are potentially expressible in such a language for intuitive theories. This suggests the hypothesis that the causal Bayes network framework 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. 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 (Gelman, Carlin, Stern, & Rubin, 1995), and Bayesian inference over such a model describes an ideal learner of abstract knowledge (Tenenbaum, Griffiths, & Kemp, 2006). 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. Kemp, Perfors, & Tenenbaum, 2007) 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 hierarchical Bayesian model 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 (Kemp, Goodman, & Tenenbaum, 2007), and simple relational theories (Kemp et al., 2008). 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 hierarchical Bayesian model, 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 Carey (2009) that part of infants' core knowledge is in the form of *perceptual input analyzers*: modules that perform simple transformations of raw perceptual input, making it suitable for conceptual cognition. These perceptual input analyzers may 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 Saxe & Carey, 2006), could be an analyzer which highlights events resulting from one's own actions, making the latent concept of intervention more salient. Alternatively, an innate or early-developing agency-detector might help in identifying interventions resulting from the actions of intentional agents. 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.

The ideal learning results that we describe below have implications for current work on the development of causal understanding and they have more general implications for the debate surrounding innate knowledge. There is a long history of philosophical speculation about the origin of the abstract causal sense, including those who thought that this must be an innate component of cognition (e.g. Hume) and those that thought it could be constructed from more concrete starting points (e.g. Maine de Biran, Michotte). More recently empirical results have shown that aspects of causal knowledge are present from early infancy, but have given little evidence that a full notion of cause is innate (Saxe & Carey, 2006). Indeed, very recent empirical results suggest that some aspects of the adult causal sense are not available for children as old as 18 months (Meltzoff, 2007), or even 24 months (Bonawitz et al., 2010). Despite this philosophical and empirical interest, there have been no computational investigations into the learnability of abstract knowledge of causality, nor what learning dynamics may emerge from the interaction of representational abilities and different sources of evidence. In the following sections we first formalize aspects of the causal Bayes network framework 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. In the Discussion we consider the implications of our results both for empirical investigation into the origin of the causal sense, and for core theoretical questions of cognitive development more generally—What must be innate? What can be learned from different kinds of input? What knowledge must be present to enable learning of other knowledge?

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 (Pearl, 2000) describes the structure of dependence between events, isolating a special role for a set of interventions. Causal Bayes

[1] Cf. the "curse of dimensionality," which describes the exponential growth of the space of possible hypotheses as the number of dimensions grows.

Law #1:	$\forall x\, \forall y\, A(x) \rightarrow \neg R(y,x)$	A-variables are exogenous.
Law #2:	$\forall x\, A(x) \rightarrow \exists y\, R(x,y)$	A-variables have at most one child.
Law #3:	$\forall x\, F_1(x) \rightarrow A(x)$	Fetaure 1 is diagnostic for A-variables.
Law #4:	$\forall x\, F_2(x) \rightarrow A(x)$	Fetaure 2 is diagnostic for A-variables.
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:	$\forall x\, \forall y\, A(x) \rightarrow \neg R(x,y)$	A-variables have no children.
Law #11:	$\forall x\, \exists y\, \neg A(y) \wedge R(y,x)$	Variables have at most one parent that is not an A-variable.

Table 1
Eleven laws that can be expressed in the language for theories. The predicate A is initially meaningless; given laws 1 and 2 it becomes the causal Bayes network notion of intervention.

networks (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) For each causal system the intervention set is known. 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 the causal Bayes network theory and a wide set of alternative theories, and should build these theories by combining simple primitive units. Kemp et al. (2008) 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 can play a 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, however in the causal Bayes network theory it will play the role of defining intervention. Finally, the two predicates, $F_i(\cdot)$, are observable features of variables. These can be thought of as perceptual features of events, extracted by dedicated input analyzers[2]. These perceptual features are meant to represent both features that could be very useful in bootstrapping causality—such as Michottean launching percepts or a feeling of self-efficacy—and "distractor" features that would not be useful in a general theory of causality—such as "happened in the morning."

This language can express a variety of theories that range from reasonable theories of causation, through incomplete theories, to theories that are entirely wrong for causation. It is useful to view a theory in this language as a collection (conjunction) of laws; table 1 gives examples of laws that can be expressed in this language. These include laws that express parts of the correct theory of causation (e.g. Law # 1: certain variables are exogenous), laws which are reasonable but not appropriate for causation (e.g. Law # 5: each pair of variables is directly related), and laws which do not seem very useful (e.g. Law # 6: no causal relations exist). Importantly, these laws are sufficient to capture the causal Bayes network theory of causality: (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. In addition, a variety of plausible variants can be expressed, describing alternative restrictions on dependency. Many of these theories may be useful for other domains of knowledge (e.g. social relations) though not for causation—in the simulations which follow we explore whether an ideal learner could construct a useful theory of causality from this domain-general language for theories.

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. 1).

[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).

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 Kemp et al., 2008) 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 (Kemp et al., 2008), 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. 1 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:

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

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:

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

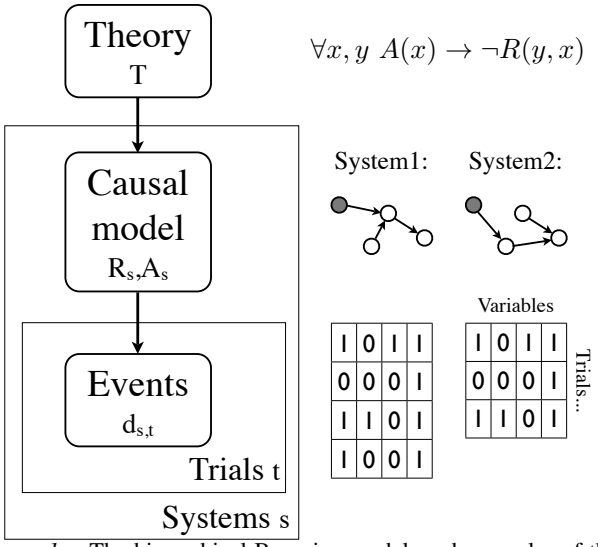

Figure 1. The hierarchical Bayesian model, and examples of the information at each level. The causal dependence relation $R(\cdot,\cdot)$ is shown as directed edges between variables (circles), the latent predicate $A(\cdot)$ is shown as shading of the variables. Binary events for each system, trial, and variable are shown as contingency tables.

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[3].

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 Table 1, 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 an intervention, then generating a causal model for each system (consistent with the correct, causal Bayes network, theory

[3] An implementation of the model can be found at: `http://www.mit.edu/~ndg/GoodmanEtAl_LTBC.tgz`

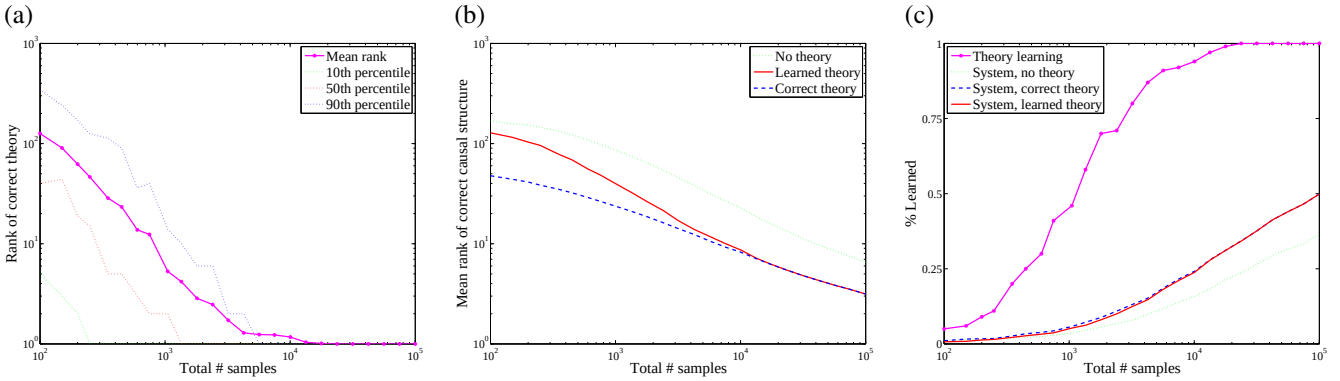

Figure 2. (a) Rank of the correct theory, mean and 10th/90th percentiles across 100 model runs. (b) Rank of the correct causal structure (mean over systems and runs), given no theory, fixed correct theory, and simultaneously learned theory. Learning abstract knowledge always helps relative to not having 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.)

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. 2a). 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. 2b 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.

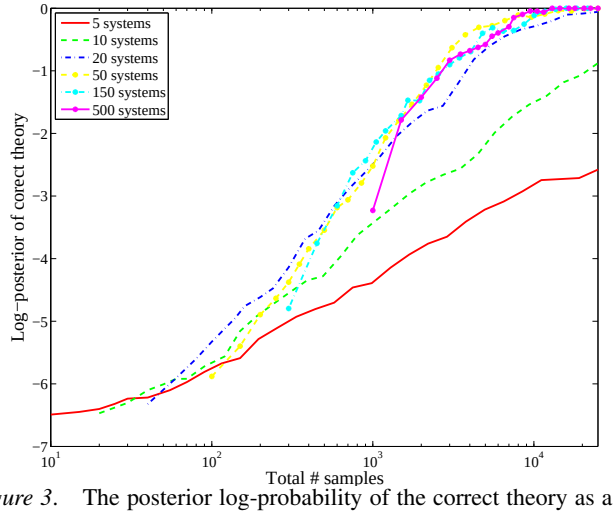

Figure 3. The posterior log-probability of the correct theory as a function of total number of samples across systems, for different numbers of systems. Each curve starts at 2 samples per system. Learning is best when evidence is gathered from many systems, even when only a few samples are taken in each system.

How can abstract knowledge appropriately bias specific learning, when it must be learned itself? Comparing Fig. 2a to Fig. 2b suggests that the correct theory is learned before most of the correct causal structures. In Fig. 2c 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 abstract 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

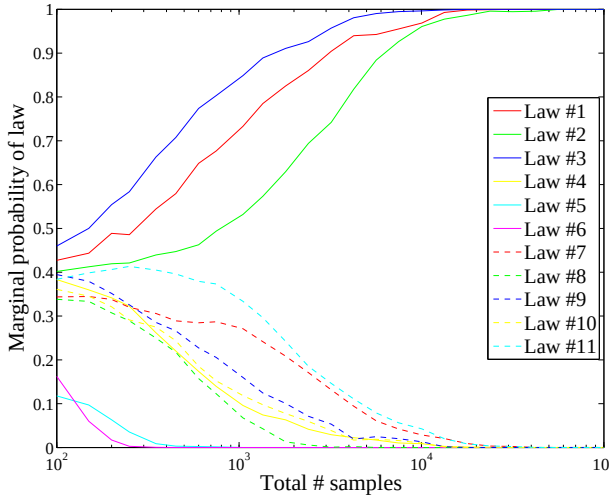

Figure 4. Learning curves for the eleven laws of Table 1.

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. 3 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. 4 shows the marginal probability of each of the eleven laws in Table 1. 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 causal Bayes networks. 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. 4, 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* (Carey, 2009): 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, proprioceptive processing as suggested by White (2009), or, more broadly, an agency-detector able to identify the actions of intentional agents (see Saxe & Carey, 2006). 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.

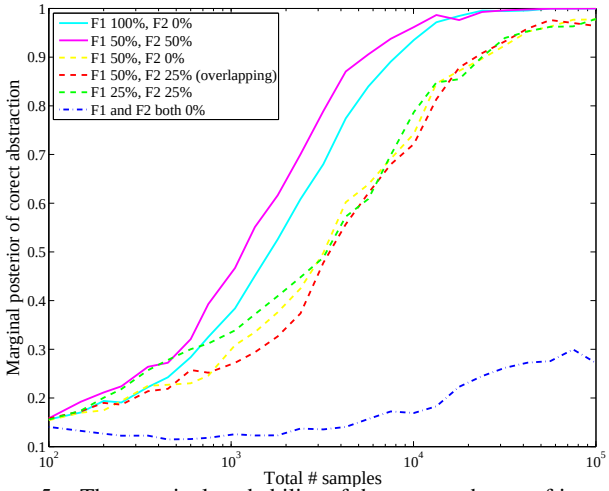

Figure 5. The marginal probability of the correct theory of intervention (i.e. Laws #1 and #2) given different sets of "input analyzers": each condition has two features which are diagnostic of intervention variables to the extent indicated (e.g. "F1 50%" indicates that the first feature covers half 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. It should be noted that each of these features could potentially be incorporated into the theory. Thus we are investigating the ability of the ideal learner to leverage partially useful features, to reject

irrelevant features, and to learn when there are no relevant features. In Fig. 5 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 relatively little evidence, often becoming entrenched before specific causal models are 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—what we termed the *blessing of abstraction*. However, in our setting the blessing of abstraction itself relied on a set of observable event features that served to make the latent concept of intervention more salient. We close by considering the significance of these results in a broader psychological context.

Given that our interest is ultimately in matters of psychological fact, not philosophical speculation, it is reasonable to ask how an ideal learnability analysis informs the study of human learning. Should it be surprising or impressive that an ideal Bayesian learner can learn a theory of causality, given data sampled from this theory? Though the mere possibility of learning may not be surprising, there are several ways that our results go beyond a demonstration of mere learnability to challenge conventional thinking about cognitive development.

First, it is often assumed that an abstract understanding of causality is innate—either necessarily or because it would be very useful—and even that this innate causal sense is what separates us from other species (Gopnik et al., 2004; Cheng, 1997). Others have suggested that the adult causal sense could be bootstrapped from earlier and more specific knowledge (Saxe & Carey, 2006), but no formal account of how this bootstrapping could work has been provided. The lack of a concrete and computationally viable learning account has likely been one of the reasons why many—even advocates of domain-general learning mechanisms—have assumed that the causal sense is largely innate. Here we have provided such a formal account, showing how the causal sense can be represented as one hypothesis in a language that also generates a broad range of alternative hypotheses, and how learning in this language can be bootstrapped from simpler structures. This work required several technical innovations, in specifying how to represent a theory of causality as a set of laws in first-order logic, how these logical laws can generate priors over graphical models, and how Bayesian inference can effectively operate over both these levels of representation to work backwards from observable data to inferences about the abstract theory underlying them.

Second, it is often assumed that abstract knowledge can only be built up slowly, or at least much more slowly than more specific knowledge. Work on the development of abstract representations in neural networks has been designed around this intuition, building up increasingly abstract layers of representation on top of lower-level, more specific layers, which are formed earlier in learning (Hinton, Osindero, & Teh, 2006). The blessing of abstraction suggests that this is not a necessary order for the construction of knowledge, but that abstract knowledge can become available before specific knowledge in any of the systems that it depends on. 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. Indeed, an abstract, domain-general theory of causality may be learned remarkably early because evidence for it may be collected from almost every experience. In future work we plan to investigate further the effects of distribution and variety of evidence; it will also be important to understand how diversity of evidence interacts with noise in the evidence, a factor we have not yet explored.

We expect that our approach to analyzing the learnability of abstract causal knowledge will be relevant for understanding the origins of abstract knowledge more generally. Whenever young infants behave as if they have some piece of abstract knowledge, it is tempting to conclude that this knowledge is innate, particularly when the abstract knowledge is present before relevant specific knowledge. This tendency may misguide—we have shown that abstract knowledge of causality can be learned so quickly that it might seem to be innate, and effectively function as an innate constraint guiding learning of more specific causal knowledge. More general versions of the framework we have described here could be applied to evaluate learnability and learning dynamics for other domain theories that have been argued to be innate—for instance the physics of objects (Spelke, 1998), or the psychology of intentional agents (Gergely & Csibra, 2003). Where innate structure *is* required to explain complex cognition, it is often assumed to be abstract conceptual knowledge (Carey, 2009). This step should also be approached with care—simpler innate structures, without conceptual content, may be sufficient when paired with a powerful learning mechanism. Finally, in domains of cognition

where abstract knowledge is clearly constructed, such as intuitive biology, it has been observed that the most abstract domain knowledge often comes into place first, before specific knowledge (Wellman & Gelman, 1998). The blessing of abstraction provides a potential explanation of this observation as well.

Though we have argued that abstract knowledge about causality may be learnable, our results should also not be taken to support an entirely empiricist viewpoint. Our ideal learner possesses 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). Our simulations suggested that these analyzers need not be perfectly tuned to causality or cover all intervention events. There are a number of plausible candidates that have been previously suggested to support causal reasoning: animacy or agency detectors, Michottean event detectors, proprioception, etc. Since a powerful learning mechanism is present in human cognition, the most efficient route to abstract knowledge may 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 input analyzers.

While our ideal learning results are purely computational, they should provide a useful viewpoint for guiding and interpreting empirical research on cognitive development. Our analysis depended on two kinds of innate (or at least preexisting) capacities that might specifically promote learning of causality, and empirical work can probe the nature, existence and development of both.

First, the language we used to represent alternative theories biases the learner by expressing some theories more simply than others. In terms of Bayesian learning, the representation language provides a hypothesis space of possible theories, and a prior based on a complexity measure over these hypotheses, much as universal grammar has been suggested to provide a hypothesis space and prior for the acquisition of natural language. A crucial difference between our language for theories and universal grammar is that our language of thought is a domain-general representational resource—out of it may be learned many theories for other domains, most of which would be useless as theories of causality. (In contrast, all grammars consistent with UG are grammars for possible human languages and only for natural languages—they don't provide theories of chemistry or causality. Griffiths and Tenenbaum (2009) have recently suggested an analogy between causal knowledge of a more specific sort and universal grammar in this narrower sense.) If causality is in fact constructed from domain-general representational resources, then a crucial project for understanding the "innateness" of causality is to characterize this "language of thought." Fortunately, the very domain generality of this resource implies that it can be studied in older children, by exploring what theories can be learned most easily in novel domains. We have begun such studies with adults (Kemp et al., 2008) and hope to extend them soon to children.

Second, perceptual input analyzers provided crucial evidence for bootstrapping the abstract theory of causality in our simulations. Traditionally, perceptual processing specific to one kind of knowledge (e.g. Michottean percepts) were seen as an alternative to abstract knowledge. Instead, we view them as a noisy signal that supports the processing and, crucially, the learning of more abstract knowledge. Understanding what is innate or learned about causality may thus depend crucially on empirically characterizing the lower-level perceptual mechanisms that respond preferentially to causal stimuli, and how these mechanisms contribute to early abstract learning.

Finally, a key empirical question that must be answered is whether children ever consider alternative theories or frameworks for causality en route to the adult causal sense. Meltzoff (2007) suggests that children's developing causal sense can be framed in terms of two alternative views outlined by Jim Woodward: Infants initially adopt an "agent causal view," in which they appreciate effects of their own actions and recognize that these relationships also apply to the actions of other people, or other entities that can be recognized as agents. Only around 18 months do infants come to a "fully causal view" in which they truly grasp what an intervention is: that the same causal relationships they intervene on can also be used by other agents and can exist in the world independent of any agent. This account is broadly consistent with the trajectory followed in our ideal learning analysis if we assume that the learner has access to input analyzers including both self-efficacy cues and (perhaps less reliably) cues about the actions of other agents. The trajectories we have observed suggest that further experiments should aim to tease apart the availability and developmental trajectory of abstract knowledge by young children. For instance, studies might investigate when children become able to use particular features to identify an intervention in a novel causal system by testing whether they use this intervention to de-confound evidence that is otherwise ambiguous between several causal structures. Recent experiments by Bonawitz et al. (2010) use a similar method to show that as late as 24 months, children's ability to appreciate candidate interventions in a novel causal system is enhanced by agent cues as well as two other sources of information: physical contact between cause and effect objects, and event descriptions using causal language. Language and culture represent an enormously important source of evidence that children have about the world, and particularly about abstract relations that may not be directly observable from sense data. Both Meltzoff (2007) and Bonawitz et al. (2010) suggest that causal language may be a useful cue to causal structure, and may even be partly responsible for the final theory of causality that children achieve. However, as we have shown, the blessing of abstraction, supported by much earlier cues to interventions such as self-efficacy and agency detection, likely

implies that some abstract knowledge of causality will be in play long before children have begun to learn language. This early abstract causal knowledge may then serve as scaffolding for a later, linguistically mediated and more sophisticated understanding of causation.

An ideal learner analysis can tell us what learning behaviors are possible. Another investigation is needed to tell us which aspects of ideal learning are possible to implement in practice; this is especially true where knowledge structures are complex and naive approaches to learning are very inefficient, as in the model described in this paper. Fortunately, a large body of work in machine learning and statistics suggests that it is possible to efficiently approximate ideal learning by stochastic search over hypotheses (MacKay, 2003). In recent work Ullman, Goodman, and Tenenbaum (2010) have shown that stochastic search methods can provide a practical and psychologically plausible means of constructing abstract theories. This demonstrates, at least, that the behaviors of an ideal learner can also be found in an efficient learner.

References

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., et al. (2010). Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., & Feldman, J. (2007). Compositionality in rational analysis: Grammar-based induction for concept learning. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for bayesian cognitive science.* Oxford: Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32. Available from `http://dx.doi.org/10.1037/0033-295X.111.1.3`

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory based causal induction. *Psychological Review*.

Hinton, G., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hume, D. (1748). *An enquiry concerning human understanding.*

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the twenty-ninth annual meeting of the cognitive science society.*

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

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

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge Univ Pr.

Meltzoff, A. N. (2007). Infants' causal learning: Intervention, observation, imitation. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford University Press, USA.

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

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, *123*(1-2), 144–165.

Spelke, E. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, *21*, 181–200.

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

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Ullman, T., Goodman, N., & Tenenbaum, J. (2010). Theory Acquisition as Stochastic Search. In *Proceedings of thirty second annual meeting of the cognitive science society.*

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society* (p. 1102-1107). Mahwah, NJ: Erlbaum.

Wellman, H., & Gelman, S. (1998). Knowledge acquisition in foundational domains. In D. Kuhn & R. S. Siegler (Eds.), (Vol. 2, pp. 523–573). Wiley.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological review*, *116*(3), 580–601.

Woodward, J. (2003). *Making things happen : a theory of causal explanation*. New York: Oxford University Press.