Cause and Intent: Social Reasoning in Causal Learning

Noah D. Goodman, Chris L. Baker, Joshua B. Tenenbaum
{ndg, clbaker, jbt}@mit.edu
MIT, Dept. of Brain and Cognitive Sciences

Abstract

The acquisition of causal knowledge is a primary goal of childhood; yet most of this knowledge is known already to adults. We argue that causal learning which leverages social reasoning is a rapid and important route to knowledge. We present a computational model integrating knowledge about causality with knowledge about intentional agency, but using a domain-general mechanism for reasoning. Inference in this model predicts qualitatively different learning than an equivalent model based on causality alone or a hybrid causal-encoding model. We test these predictions experimentally with adult participants, and discuss the relation of these results to the developmental phenomenon of over-imitation.

Keywords: Causal learning, social cognition, Bayesian modeling, imitation.

Introduction

How do children acquire conceptual knowledge? One answer is that children are adept at rational inference from direct experience with the world—children as scientists (Gopnik, Meltzoff, & Kuhl, 1999). Human culture suggests another answer: the quickest route to conceptual knowledge may be by learning what others already know (Tomasello, 1999; Gergely & Csibra, 2006). Indeed, a vast majority of the knowledge that a child will acquire is already known by adults in their society. This suggests that children come equipped with social learning mechanisms to encode the knowledge of adults (Lyons, Young, & Keil, 2007). We suggest a middle ground between these two views: that an understanding of intentional agency makes it possible to use social context as a source of evidence to enable rapid learning without requiring dedicated psychological mechanisms.

Among the most profound achievements of human knowledge is an understanding of causal structure in the world; not coincidentally, causality has been a major area of research into children's ability to learn from evidence (Gopnik et al., 2004). In this paper we therefore focus on the acquisition of causal knowledge in a social, but non-linguistic, setting. To explore the hypothesis that social learning emerges from the interaction, under domain-general inference processes, of existing conceptual structures, we construct a computational model integrating social and causal representations. Recent modeling of human causal reasoning has focused on the causal Bayes nets approach (Pearl, 2000; Gopnik et al., 2004). This view helps explain how causal learning can succeed based on observed co-occurance between events and well-chosen interventions. The interventions, however, are treated as unexplained actions on a system. In contrast, recent models of intuitive psychology have focused on the selection of actions by agents given their goals and beliefs (Baker, Tenenbaum, & Saxe, 2007; Goodman et al., 2006), but treated the structure of the world as background knowledge available to all agents. We combine these modeling approaches by constructing a model of intuitive psychology in which beliefs about the causal structure of the world are represented and used for action selection, and showing how such an intuitive theory of causal-agency can explain the source of interventions and speed causal learning.

We test this model experimentally by studying adult intuitions in a set of scenarios that provide both social and causal information. These scenarios are chosen to distinguish between the combined social-causal model and two alternatives: a similar causal-only model, and a hybrid gated-inference model. We then use these results to explain a puzzling phenomenon of imitation-based learning in children.

Computational modeling

Our goal is to construct a formal model which simultaneously represents knowledge about causality and knowledge about intentional agency, and to explore how Bayesian inference over the combination differs from inference over each piece in isolation. Knowledge can be represented as probabilistic generative models; Bayesian inference then "inverts" this generative knowledge, specifying appropriate beliefs about latent states given observed evidence. We begin by recalling standard generative models capturing (aspects of) causality and intentional agency, then describe how they may be integrated, and finally describe predictions of the resulting model.

Causality A causal Bayes net (CBN) model describes the probability $P(E|A,S)$ of observing a set of events E, given the causal structure S, and the interventions, or exogenous actions, A (Fig. 1a). More formally, a CBN consists of a directed acyclic graph on a set of *variables*, together with a specification of the probabilistic dependence of each variable on its parents in the graph. The variables represent events or states, and the edges represent the fact of a causal dependence. For some variables there is an *intervention*: an exogenous event that forces the variable to a particular value, irrespective of the values of its parents. We will assume that the dependencies between variables are described by *noisy-or* functions (each parent is a sufficient cause) or *noisy-and* functions (parents are jointly sufficient and individually necessary); the causal strength of these dependencies, ε, is a fixed parameter. (See Pearl (2000) for more about the formalism and uses of causal Bayes nets.)

Agency Bayesian decision theory (Berger, 1985) describes the choices made by a rational agent facing a *stochastic decision problem* (SDP) (Fig. 1b). A SDP consists of a set of possible actions the agent may take, a utility function $U(E)$,

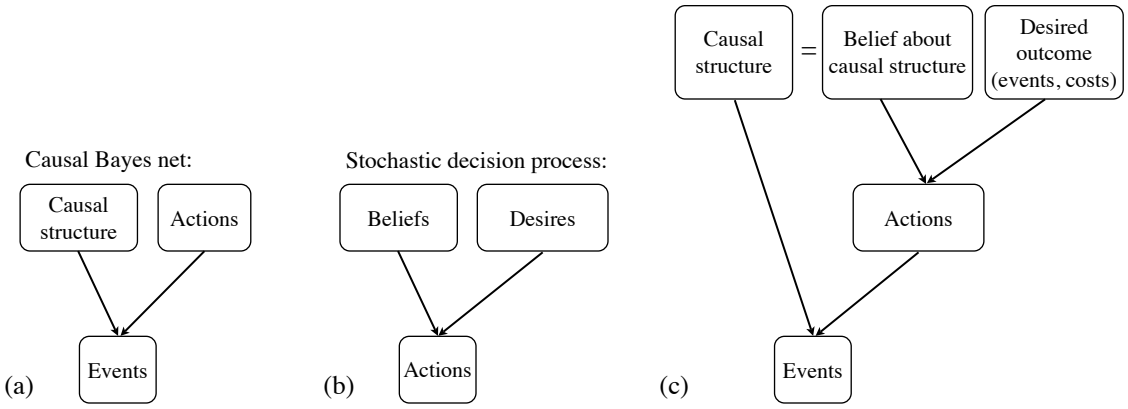

Figure 1: Schematic representations of generative models for causality (a), intentional agency (b), and the social-causal combination (c). Equality between the true causal structure and an agent's belief about causal structure embodies the knowledgeable agent assumption.

capturing the agent's desires (reward for each possible outcome E), and a belief function $P_B(E|A)$, capturing the agent's beliefs about how the world works (the likely outcomes of an action A). Bayesian decision theory specifies that the agent should choose an action to maximize her expected utility: $\mathbf{E}_{P_B(E|A)}U(E)$. If we assume that agents are only approximately rational, and hence only softly maximize expected utility:

$$P(A|U,P_B) \propto e^{\beta \cdot \mathbf{E}_{P_B(E|A)}U(E)}, \tag{1}$$

where the parameter β determines the amount of decision noise.

Eq. 1 can be used to model the intuitive theory of intentional agency of a person who makes the *rational agent assumption* (Baker et al., 2007).

Integration In order to capture reasoning about an intentional agent choosing actions based on their causal knowledge, we construct a model which integrates the above approaches to causality and agency. We first assume that the observed agent represents the world in terms of causal Bayes net S: the set of outcomes is the set of all possible events (variable values), the set of actions is all combinations of interventions, and the belief function is described by the CBN dependency: $P_B(E|A)$=$P(E|A,S)$. Further, we assume that the utility function of the agent splits into a cost, $\mathcal{C}$, for each intervention made, and a reward, $\mathcal{R}$, for each desired event achieved.[1]

This CBN-based stochastic decision problem describes a simple theory of mind for reasoning about a causal agent—the intuitive theory one person has about another person's causal knowledge, desires, and actions. Note that a person might represent the causal structure of the world via a CBN, and represent another agent's beliefs via a second CBN (wrapped inside a SDP). When the interventions that enter the person's own causal reasoning are the actions of the other agent, they need not be treated as unexplained events. Fig. 1c represents a combined model in which the interventions of the CBN have been identified with the actions of the SDP.

We will simplify by making the *knowledgeable agent assumption*: the beliefs of the observed agent about the causal structure of the world reflect the true (but unknown to the observing agent) causal structure of the world. While this is clearly not always the case, people, and especially children, are often in situations where they can observe the actions of an expert on a novel-to-the-observer causal system. In Fig. 1c this is represented with the equality between the true causal structure of the world and the observed agent's beliefs about the causal structure. (It is possible to relax this assumption, leading to a model which incorporates explicit reasoning about belief formation and update; see Goodman et al. (2006) for a related model.)

Given this setup, Bayesian inference can be used to infer a causal structure from observation of events and actions in two ways: assuming only causal knowledge (causal-only inference), and assuming both causal and social knowledge (social-causal inference). For causal-only inference, the posterior over causal structures is given by:

$$P_{\mathrm{c}}(S|A,E) \propto P(E|A,S)P(S), \tag{2}$$

where $P(S)$ is the prior probability over causal structures—we take this to be given by an independent prior probability $P(v_i \rightarrow v_j)$ that each potential edge is in S.

For social-causal inference the joint posterior over causal structure, S, and the (unknown) utility function, U, of the agent performing the actions is given by:

$$\begin{aligned} P_{\text{s-c}}(S,U|A,E) &\propto P(A,E|S,U)P(S)P(U) \\ &\propto P(E|A,S)P(A|S,U)P(S)P(U), \end{aligned} \tag{3}$$

where $P(A|S,U)$ is given by Eq. 1. (Note that the same causal structure, S, enters both the CBN term and the SDP term of Eq. 3—this is the knowledgeable agent assumption.) We assume a uniform prior on sets of desired events, which determines $P(U)$. If we are interested in the causal structure alone,

[1]The model described in this section can be easily extended to capture temporal effects by using dynamic Bayes nets and Markov decision process models.

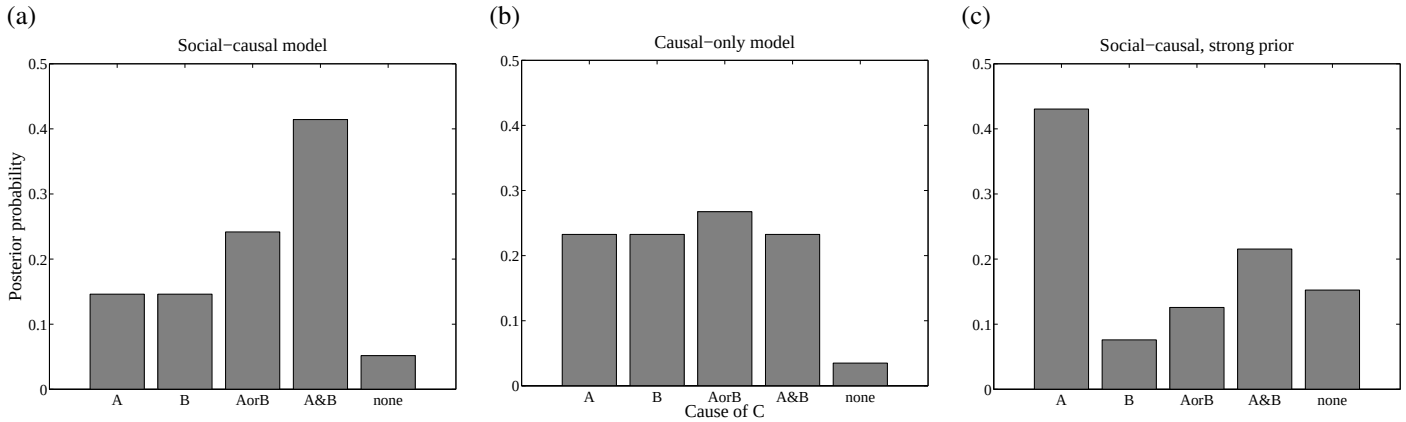

Figure 2: Model predictions: the social-causal (a) and causal-only (b) models (both with uniform causal prior: $P(\text{A}\rightarrow\text{C})=P(\text{B}\rightarrow\text{C})=0.5$), and (c) the social-causal model with prior disfavoring cause B: $P(\text{B}\rightarrow\text{C})=0.15$. (In all cases $\mathcal{R}=6$, $\mathcal{C}=1$, $\varepsilon=0.85$.)

we can marginalize over these utility functions:

$$P_{\text{s-c}}(S|A,E) = \sum_U P(S,U|A,E). \quad (4)$$

Predictions We describe the simplest scenario in which the causal-only and social-causal models make qualitatively different predictions about causal learning. Imagine a situation with three causal variables: two potential causes, A and B, and one potential effect, C. Simultaneous interventions on A and B are observed, and activation of C follows. To the causal-only model this is *confounded* evidence, and it is unable to distinguish possible causal relations[2] (Fig. 2b). If we assume, however, that the simultaneous interventions on A and B are the actions of a knowledgeable agent, the social-causal model makes the (strong) inference that *both* A and B are required to bring about C (Fig. 2a). This inference follows from the tradeoff of costs and goals in the social-causal model. Informally: the agent wishes to minimize action cost while achieving a desired outcome, because each interventions has a cost, the most parsimonious inference is that the agent believes both actions are required to bring about her desired outcome—if her goal is to bring about C, this means that she must believe the causal structure is A&B→C.

If the prior probability of cause B is significantly less than that of cause A (i.e. $P(\text{B}\rightarrow\text{C})\ll P(\text{A}\rightarrow\text{C})$) the social-causal inference model overrides the social inference, instead concluding that A is the only cause of C; Fig. 2c. However, prior beliefs integrate continuously, coming to dominate the inference when they are fairly strong, but influencing inference even when they are weak (Fig. 3). This graded behavior contrasts with another possible mechanism for incorporating causal knowledge, the *gated-encoding model*, in which social context is used for inference, but prior beliefs serve as a gating mechanism, forming the "boundary conditions" for attention to social context (Lyons et al., 2007).

[2]The causal-only model exhibits a slight preference for the structure AorB→C because the evidence is most likely given this structure: either of the events is a sufficient cause of C.

The social-causal inference model has a number of free parameters—the causal strength ε, the action cost $\mathcal{C}$, the goal reward $\mathcal{R}$, and the decision noise β—that affect quantitative predictions. However, the qualitative predictions seen in Figs. 2 and 3 hold over a wide range of parameters. These predictions differ from the predictions of a causal-only inference model (which is unable to use social information to deconfound ambiguous evidence), and a gated-encoding model (which fails to continuously integrate prior causal knowledge with social information).

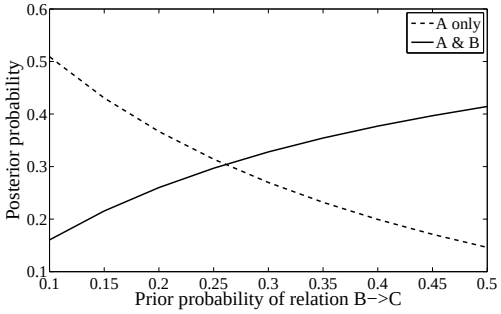

Figure 3: The graded effect of prior knowledge on inferences of the social-causal model.

Experiment: Causal learning in social context

In the following experiment we test the qualitative predictions of the social-causal model that critically distinguish it from the other possible models: that people will use social context as a source of information to disambiguate confounded causal evidence (Fig. 2a), that this relies on a knowledgeable agent assumption (Cf. Fig. 2b), and that this interacts (in a graded fashion) with prior causal knowledge (Figs. 2c and 3). We constructed scenarios in which a knowledgeable agent performed two actions (simultaneously) and an effect followed (the Social condition). To show that inferences follow from rational- and knowledgeable-agent assumptions, and not extraneous non-social factors of the scenarios, we controlled

for mere agency by manipulating the knowledgeability of the actor about the causal system—in the Self condition we described the actions as being taken by "you" (the participant). Finally, to explore the effect of prior causal knowledge we constructed a third variant of the scenarios (the Prior condition) in which one of the two actions was relatively implausible as a cause of the effect. To verify that these causes were implausible we followed the main experiment by eliciting plausibility judgments for each potential causal relation.

Method

Participants Participants were fifteen members of the MIT community who received a small compensation for their time.

Materials We constructed a series of scenarios based on the abstract causal scenario described above. Each scenario consisted of three sentences: (1) setup of scenario, (2) agent/you performs two actions simultaneously, (3) effect follows. Each scenario had three variations differing only in the second sentence: in the Social condition the agent performed two equivalent actions, in the Self condition "you" (the reader) perform two equivalent actions, in the Prior condition the agent performs two dissimilar actions, one of which is an implausible cause of the effect. In order to rule out the hypothesis that social reasoning effects are peculiar to a single domain (e.g. artifacts), we constructed scenarios drawn from three different domains: mechanical (artifact), biological, and chemical. In each domain we constructed three different scenarios, for a total of nine scenarios. Each of the nine scenarios had three variants: one for each of the three conditions.

For example, one Social condition scenario in the biological domain was:

> You work at a genetically-engineered plants nursery, and one of your coworkers is tending to some almost-dead flowers that you havent seen before. Your coworker simultaneously pours a yellow liquid and a blue liquid on the flowers. By the end of the day, the flowers are growing again.

In the Self condition the middle sentence was changed to:

> One day when your coworker is gone, you find a yellow liquid and a blue liquid in his supplies and simultaneously pour them on the flowers.

In the Prior condition the middle sentence was changed to:

> Your coworker simultaneously drinks a yellow liquid and pours a blue liquid on the flowers.

Procedure Participants were assigned to conditions such that each participant received one scenario in each condition in each domain (assignments were counterbalanced in a latin square design). The order of scenarios was randomized between participants. After reading each scenario participants were asked for their causal structure inferences ("What causes C?") in the form of bets (which were required to sum to \$100—hence, providing a natural elicitation of probability judgments). The five options were of the form:

> A (and not B)
> B (and not A)
> Either A or B (or both together)
> Both A and B together (but not either one alone)
> C is unrelated to A and B,

with A, B, and C were replaced with the appropriate events. The order of the response options was consistent for each participant, but randomized between participants.[3] Following the main portion of the experiment participants were asked to rate the plausibility (on a seven-point scale) of each action causing the corresponding effect.

Results and discussion

Fifteen (out of 135) responses failed to sum to 100 (never by more than 10); these responses were normalized to 100. Ratings on the plausibility check were left blank on two responses; these were omitted from our analyses.

In no condition was there a significant effect of domain; we collapse across domains for the remaining analyses. Fig. 4 summarizes the causal structure inferences of participants in each of the three conditions.

Consistent with the predictions of the social-causal model (Fig. 2a), bets placed on "A and B" in the Social condition were significantly greater than bets placed on any other option (vs. "A only": $t(44)=8.58$, $p<0.001$; vs. "B only": $t(44)=8.59$, $p<0.001$; vs. "A or B": $t(44)=3.25$, $p<0.01$; vs. "no relation": $t(44)=8.47$, $p<0.001$). (All t-tests are two-tailed and, where appropriate, correlated-samples.) Thus, in contrast to the predictions of a causal-only learning model, participants infered that A and B together cause C, despite confounded evidence. To verify that this inference was based on social context information, we compare the Social condition to the Self condition, in which the knowledgeable agent assumption should be weakened. Indeed, the "A and B" bets were significantly less in the Self condition than in the Social condition ($t(44)=3.34$, $p<0.001$), and in the Self condition there was no longer a significant difference between "A and B" and "A or B" responses ($t(44)=1.71$, $p=0.094$)[4].

The prior plausibility check confirmed that the causal relations intended to be implausible were significantly less plausible than those intended to be plausible ($t(44)=19.27$, $p<0.001$). As predicted (Fig. 2c), prior plausibility affected causal structure inferences: the bets on "A and B" were significantly greater in the Social condition than in the Prior condition ($t(44)=5.11$, $p<0.001$). Thus participants used prior

[3] To verify that participants were considering each scenario, we inserted an attention check at a random position within the experiment packet. This page looked visually similar to other pages but contained instructions to write only "I have read the instructions" and proceed. No participant failed this check.

[4] In the Self condition there was a trend toward "A and B" bets. Informal debriefing suggested that some participants misunderstood the fictional assumption of the scenario, treating "themselves" as knowledgeable agents. If this was the case, we would expect to observe a mixture of the causal-only model with the social-causal model; this is consistent with the observed trend.

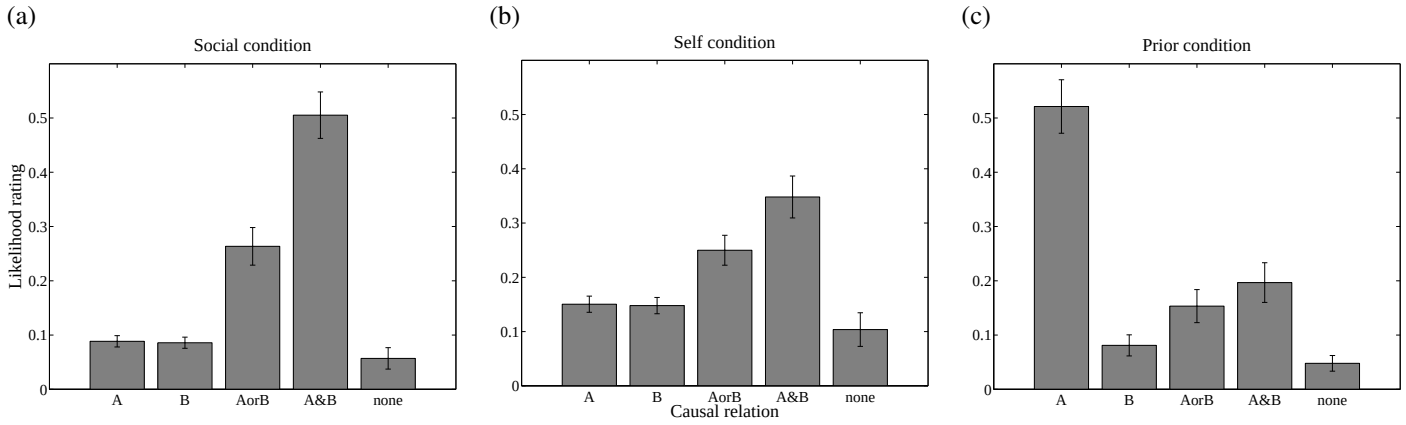

Figure 4: The mean bet (likelihood rating) placed on each of the five possible causes of C. The Social condition (a) confirms the social-causal model predictions (Fig. 2a). The Prior condition (c) confirms predictions of the social-causal model with strong prior (Fig. 2c). The Self condition (b) reflects a reversion to the causal-only model (Fig. 2b), as expected, but seems to be mixed with residual social-causal inferences—see Footnote 4.

causal knowledge to inform inferences, even when social context information was available. To see whether this is a graded integration of information sources, as predicted by the social-causal model (Fig. 3), or an all-or-nothing gating effect of prior knowledge, we exploit natural variation among the scenarios. The relationship between the plausibility rating of a participant and their bet on "A and B" in the corresponding scenario, can be used to further examine the effect of prior knowledge on inferences. Pooling Social and Prior scenarios, prior plausibility ratings explain 43% of the variance in bets (r=0.66, p<0.001), as shown in Fig. 5.[5] Within conditions, causal structure inferences remain significantly correlated with the variation in plausibility judgments (r=0.46, p<0.01 within the Social condition, r=0.56, p<0.001 within the Prior condition). This result indicates that participants continuously integrate prior causal knowledge with social context information, rather than using prior causal knowledge as a gate on social inference.

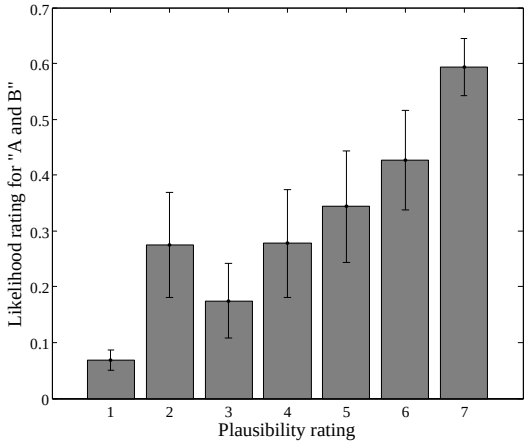

Figure 5: The mean bet (likelihood rating) of participants on "A and B" according to their plausibility rating for B as a cause of C. The graded effect of prior knowledge confirms the model predictions (Fig. 3).

Over-imitation

The results of the previous sections show that generic inference abilities, combined with an understanding of causality and agency, can result in rapid learning of causal knowledge. Yet where there is rapid learning there is the possibility of going rapidly astray—are there situations in which social-causal inference might lead to incorrect conclusions?

A number of authors have reported that children seem to over-imitate adults, copying even actions which are, to adults, clearly superfluous to bringing about an effect (Horner & Whiten, 2005; Lyons et al., 2007; Meltzoff, 1995). For instance Horner and Whiten (2005) present a "puzzle box" to children and demonstrate a series of actions which culminate in retrieving a prize from within the box. The box is transparent, and some of these actions are plausibly related to the outcome, but one is not (for example, touching a rod to the top of the box). When invited to retrieve the prize, children perform *all* the actions, including the superfluous one. Chimps in a similar experiment did not over-imitate, leaving out the implausible action. Lyons et al. (2007) investigated a number of possible explanations for over-imitation in children but found it to be remarkably robust; the only manipulation they report that reversed children's over-imitation was removal of physical contact between cause and (potential) effect (Lyons et al., 2007, Expt. 2b). On the basis of these findings Lyons et al. (2007) suggest that over-imitation reflects an "automatic causal encoding" mechanism, with "boundary conditions" to switch off this encoding (such as physical contact).

Our modeling results indicate that a separate principle (such as automatic causal encoding) needn't be invoked to explain children's over-imitation. If children's prior beliefs are weaker than adults' (and, like adults, contact-causality is

[5]The correlation is higher for group means (r=0.85); we are, however, primarily interested in the relationship within individual participants.

amongst the strongest priors), then over-imitation behavior follows as the result of domain-general probabilistic inference, assuming that children use social context as a source of evidence. Our experimental results, which demonstrate similar inferences in adults, further support this interpretation by showing that the interaction between prior causal knowledge and social inference is not a gating mechanism (as suggested by "boundary conditions" on automatic causal encoding), but a graded integration as required by Bayesian inference.

It is interesting to ask what the differences between chimps and children mean, in light of this interpretation. The causal-only model predicts little over-imitation, since even a small amount of prior bias dominates the inference. This suggests that chimps too are acting rationally on the basis of causal structure inferences, but are failing to use social information to guide these inferences. That is, the quixotic over-imitation of children may reflect a deep understanding of other minds, while the lack of this behavior in chimps might reflect poor theory of mind.

Conclusion

We have presented a computational model of the social acquisition of causal knowledge. This model integrates existing Bayesian approaches to causal reasoning and intuitive psychology. It predicts qualitatively different inferences about causal structure, given social context, than an equivalent model based on causality alone, and predicts that prior causal knowledge will be integrated into inferences in a graded fashion. We verified these predictions experimentally with adult participants, and discussed the relation of these results to the developmental phenomenon of over-imitation.

While research on imitation and social cognition has stressed the tendency of children to gain knowledge directly from adults, research on causal learning has focused on the ability of children to extract causal structure from observation and interaction directly with causal systems. Our results suggest that social context may provide a crucial, and often ignored, source of information that children use to learn causal knowledge. However, the ability to learn from disparate sources of evidence raises an important empirical question: to what extent do children actually rely on social evidence (vs. observation, exploration, etc.) in acquiring causal knowledge? Our results indicate that social learning depends on a knowledgeable-agent assumption, thus it is reasonable to start by asking when children are aware that adults have knowledge that they themselves lack. Kushnir, Wellman, and Gelman (2008) have recently shown that children are sensitive to the knowledgeability of others, and treat intentional actions by knowledgeable agents as more informative about causal structure than actions by unknowledgeable agents. This is consistent with our experimental results; our modeling results show that such inferences are an appropriate response to social context, rather than a rough heuristic, or incorrect bias.

Throughout this paper, social context consisted of observed actions of an agent with a concrete goal (e.g. reviving a flower). Different inferences might be licensed when the agent has a social goal, such as communication or pedagogy (e.g. teaching how to revive a flower—see Shafto and Goodman (2008)). Further work will be required to distinguish learning based on inference of concrete goals from that based on social goals.

We have suggested that rapid social learning follows from domain-general inference abilities and an intuitive theory of other minds. This emerged naturally in our computational model by considering a powerful inference mechanism (Bayesian inference) operating over complex knowledge structures. Since the compositionality of concepts is a crucial feature of human cognition, it is likely that other important aspects of human thought also lurk in the interactions between representations that are well understood in isolation. Studying the effects of such interaction is thus an especially important and potentially fruitful direction for computational cognitive science.

References

Baker, C., Tenenbaum, J., & Saxe, R. (2007). Goal inference as inverse planning. *Proceedings of the 29th annual meeting of the cognitive science society*.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Gergely, G., & Csibra, G. (2006). Sylvias recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. *Roots of human sociality: Culture, cognition, and human interaction, ed. NJ Enfield & SC Levenson.*

Goodman, N., Baker, C., Bonawitz, E., Mansinghka, V., Gopnik, A., Wellman, H., et al. (2006). Intuitive Theories of Mind: A Rational Approach to False Belief. *Proceedings of the Twenty-Eigth Annual Conference of the Cognitive Science Society*.

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

Gopnik, A., Meltzoff, A., & Kuhl, P. (1999). *The Scientist in the Crib: Minds, Brains, and How Children Learn.* William Morrow & Col., Inc.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition*, *8*(3), 164–181.

Kushnir, T., Wellman, H., & Gelman, S. (2008). The role of preschoolers social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–1092.

Lyons, D., Young, A., & Keil, F. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751-19756.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838–850.

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

Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual meeting of the cognitive science society.*

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.