Developmental Science

Developmental Science (2012), pp 1–12 DOI: 10.1111/j.1467-7687.2012.01135.x

PAPER

Epistemic trust: modeling children's reasoning about others' knowledge and intent

Patrick Shafto,[1] Baxter Eaves,[1] Daniel J. Navarro[2] and Amy Perfors[2]

1. *Department of Psychological and Brain Sciences, University of Louisville, USA*
2. *Department of Psychology, University of Adelaide, Australia*

Abstract

A core assumption of many theories of development is that children can learn indirectly from other people. However, indirect experience (or testimony) is not constrained to provide veridical information. As a result, if children are to capitalize on this source of knowledge, they must be able to infer who is trustworthy and who is not. How might a learner make such inferences while at the same time learning about the world? What biases, if any, might children bring to this problem? We address these questions with a computational model of epistemic trust in which learners reason about the helpfulness and knowledgeability of an informant. We show that the model captures the competencies shown by young children in four areas: (1) using informants' accuracy to infer how much to trust them; (2) using informants' recent accuracy to overcome effects of familiarity; (3) inferring trust based on consensus among informants; and (4) using information about mal-intent to decide not to trust. The model also explains developmental changes in performance between 3 and 4 years of age as a result of changing default assumptions about the helpfulness of other people.

Introduction

Children face a daunting task in learning about the world; there are an almost unlimited number of things to learn, and the time available for learning through direct experience is limited. How might they overcome this limitation? One possibility relies on the fact that children are surrounded by people, which provides an opportunity for them to learn about the world through indirect experience. Although potentially very informative, indirect experience also poses a problem: it is not constrained to be veridical. Children must therefore be able to infer which information and informants are trustworthy if they are to take advantage of indirect experience. This is the problem of epistemic trust (Mascaro & Sperber, 2009; Pasquini, Corriveau, Koenig & Harris, 2007; Corriveau, Meints & Harris, 2009b; Corriveau, Fusaro & Harris, 2009a; Corriveau & Harris, 2009a, 2009b; Koenig & Harris, 2005a, 2005b; Clement, Koenig & Harris, 2004; Harris & Corriveau, in press; Harris, 2007; Jaswal, Croft, Setia & Cole, 2010; Jaswal & Neely, 2006).

Even preschool children can make sensible inferences about who to trust. Koenig and Harris (2005a) demonstrated that 4 year-old children can distinguish between accurate and inaccurate informants and can use this information to aid in the selection of a more accurate informant. Indeed, by 4 years of age children are quite sophisticated in their ability to monitor who to trust. For instance, Pasquini *et al.* (2007) systematically manipulated the relative accuracy of two informants in a labeling task, and found that children preferred to ask the more accurate informant about a label for a novel object (see Figure 2). Four year-olds can also use familiarity in judging informants, favoring informants with whom they have established a strong history of accuracy by default but switching to an unfamiliar informant if faced with evidence that the unfamiliar informant is more accurate (Corriveau & Harris, 2009a). In addition, children use information about consensus, trusting informants who label objects in the same way as the majority of others (Corriveau *et al.*, 2009a). These studies provide strong evidence that children use a variety of information to guide who to trust.

Researchers have interpreted these results as an indicator of children's abilities to infer which individuals are knowledgeable. Under this interpretation, individuals who label familiar objects correctly are inferred to be knowledgeable; whereas, individuals who label familiar objects incorrectly are inferred to be not knowledgeable. However, awareness of an informant's knowledge is not enough to justify epistemic trust: it is also important to understand their *intent*–whether they are trying to be helpful or deceptive. A knowledgeable but deceptive informant should be expected to consistently provide unreliable information, and should not be trusted. Given

Address for correspondence: Patrick Shafto, 317 Life Sciences, University of Louisville, Louisville, KY 40292, USA; e-mail: p.shafto@louisville.edu

this observation, one might reasonably ask, are children's abilities strictly due to inferences about knowledge, or do inferences about intent affect their judgments about who to trust?

A separate line of experiments suggests that preschool children do use information about intent to guide inference. Mascaro and Sperber (2009) presented children with a situation in which there were two different colored boxes, under one of which was hidden a candy. An informant then looked under both boxes, so the child knew that the informant knew where the candy was. In the test condition, a puppet entered and warned the child that the informant was a 'big liar' who always told lies. The informant then labeled one box, saying for example, 'The sweet is in the red box'. Their results showed that by 4 years-old children used information about the informant's intent, showing a significant preference for the cup that the informant did not indicate. This study suggests that preschool children use intent, when it is clearly marked, to guide inference.

Do inferences about intent play a role in children's epistemic trust? While researchers have interpreted previous results as indicators of children's inferences about knowledgeability, evidence suggests that around the same ages children are developing the ability to use an informant's intent to guide inference. We propose a theoretical framework within which we can investigate whether inferences about epistemic trust are due to knowledge alone (cf. Bovens & Hartmann, 2003) or whether intent plays a substantive role in children's judgments. The framework demonstrates how a learner might integrate inferences about others' knowledge and intent, while at the same time learning new labels. We formalize this framework as a probabilistic model and provide evidence that joint inference about informants' knowledge and intent provides an accurate account of 4 year-olds' behavior and explains developmental changes between the ages of 3 and 4 as a consequence of changing prior expectations about others' intent.

We proceed by introducing the model, which formalizes the relationship between evidence, inferences about knowledge and intent, and learning. Next, we demonstrate that our model of trust captures 4 year-olds' abilities but a simpler model based on knowledge alone does not. We then apply the model to 3 year-olds' performance, contrast the model's explanations for 3 and 4 year-olds' behavior, and conclude by discussing implications for learning and development.

A model of epistemic trust

Formalizing the role of epistemic trust in learning requires specifying two inference problems. First, how would a learner expect an informant to choose information to provide, and how would that depend on the informant's knowledge and intent? Second, how would the learner use the information provided by the informant to simultaneously make inferences about the true state of the world and about whether to trust the informant? We present a model that unifies these problems under a single framework, and provides an account of how children may simultaneously learn about the world and whether to trust an informant.

Because the studies we model involve learning the correspondence between objects and their labels,[1] the 'true state of the world' we seek to model is the set of correct labels in a word-learning task. We adopt a probabilistic modeling framework in which learning is based on data and formalized as Bayesian inference (Tenenbaum, Griffths & Kemp, 2006). In Bayesian inference, a learner's beliefs after observing some data (their posterior beliefs) are related to their prior beliefs as well as how well those beliefs explain the data (the likelihood). In this case, the beliefs we seek to model include children's beliefs about their world (i.e. the correct label or labels for an object or objects) and their informant (i.e. how knowledgeable and helpful the informant is). It is necessary to specify prior beliefs about each of these characteristics. The other key component of our approach is the likelihood–specifically the sampling model underlying the calculation of the likelihood. Precisely which label an informant provides depends on their knowledge and intent, as well as the true label; a learner, after observing the labels, can therefore reason backwards about all three of these things, given certain (sampling) assumptions about how knowledge and intent translate into the choice of a label.

In our model, epistemic trust is assumed to depend on the knowledgeability of the informant (denoted k) as well as the extent to which he or she intends to be helpful (denoted h). If we let l denote the actual label that the informant provides, then the goal of the learner is to infer the most likely state of the world s (i.e. the correct label for an object) and nature of the informant (k and h) given that label. Formally, this corresponds to calculating $P(s,k,h|l)$, which according to Bayes' rule is given by:

$$P(s,k,h|l) \propto P(l|s,k,h)P(s,k,h), \qquad (1)$$

where $P(s,k,h) = P(s)P(k)P(h)$ assuming the prior probability of s, k, and h are independent of one another. Note that s and l correspond to possible object labels: s is the correct label and l is the label given to the learner by the informant. The two 'social' characteristics are binary: the informant is either knowledgeable ($k = 1$) or not ($k = 0$), and is either trying to help ($h = 1$) or trying to hinder ($h = -1$) (cf. Ullman, Baker, Macindoe, Evans, Goodman & Tenenbaum, 2010). However, learners' beliefs *about* these characteristics are distributions over the possible values that they can take.

Viewed in this fashion, it becomes clear that different experimental manipulations correspond to different prior assumptions. For instance, if a child observes a new

[1] In one case, what is being learned is an object's location, but, as we will argue, the formal problem is the same.

informant labelling an object whose label is already known to the child, then $P(s)$ is a point mass distribution which assigns probability 1 to the correct label and probability 0 to all other labels. When this happens, the learner can leverage their knowledge about the true state of the world to draw inferences about the informant's knowledge k and helpfulness h. Alternatively, the child might not know the true label; in this case, if there are n possible labels, it would be sensible to assume, as our model does, that $P(s) = 1/n$ for all possible states of the world (i.e., all possible labels). If the child is provided with a label from an informant whose knowledgeability and/or helpfulness is known from past experience, then the priors $P(k)$ and $P(h)$ can be adjusted to capture this information. In more complex situations where labels are provided (for unknown objects) by multiple informants whose knowledge and helpfulness are not known, the learner must simultaneously infer the correct label s, along with the knowledge k_i and helpful intent h_i of each informant i.

All of the scenarios described above can be captured by the Bayesian learning rule in Equation 1, in which the link between prior beliefs $P(s,k,h)$ and posterior beliefs $P(s,k,h|l)$ is supplied by the likelihood function $P(l|s,k,h)$. The likelihood function, in this case, reflects the learner's theory of how the informant would have generated a label l, if the true state of the world was s, and the informant had knowledge level k and helpfulness h. As Figure 1 illustrates, the behavior of the informant also depends on a hidden variable, b, which corresponds to the informant's belief about what the true label is, where this belief depends on how knowledgeable the informant is as well as on the true label. We express this via a distribution over beliefs, $P(b|k,s)$. Then, for any given belief b, the informant will generate the label in a manner that depends on how helpful they are, expressed by the distribution $P(s|b,h)$. Because the learner cannot directly observe the beliefs b of any informant, he or she must (in effect) average over his uncertainty about what the informant really believes in order to calculate the likelihood of seeing that label l. This is captured by the following equation:

$$P(l|s,k,h) = \sum_b P(l|b,h)P(b|k,s). \tag{2}$$

In order to complete the model, we need to specify these two distributions–one over the possible beliefs of the informant, and the other over what labels the informant might provide as a result of these beliefs. For the distribution over possible beliefs, we assume that:

$$P(b|k,s) = \begin{cases} 1 & \text{if } k = 1 \textit{ and } b = s \\ 0 & \text{if } k = 1 \text{ and } b \neq s \\ 1/n & \text{if } k = 0. \end{cases} \tag{3}$$

That is, we assume that a knowledgeable informant always has the correct belief, whereas a non-knowledgeable informant believes something chosen randomly from the set of possible labels. This is a simplification, of course, but it is sufficient for the current purposes.

The subtle aspect to the model lies in the choice of $P(l|b,h)$, the probability that an informant would use the label l given that the informant believes the true label to be b and has degree of helpfulness h. The model assumes that a helpful informant ($h = 1$) will try to select the label that maximizes the extent to which the *learner* comes to share the same belief as the informant, whereas an unhelpful informant ($h = -1$) will try to minimize this.

Formally, this is accomplished by choosing a distribution over labels $P(l|b,h)$ that satisfies the following 'communicative sampling' relationship (Shafto & Goodman 2008; Shafto, Goodman, & Griffths, under revision). In communicative sampling, the key idea is that the speaker actively seeks to shape the beliefs of the listener in a manner governed by a 'helpfulness' parameter h, and both parties are assumed to be Bayesian reasoners. Formally this means that the probability that an informant (or speaker) with belief b chooses a label l is closely related to the probability that the learner will come to share the informant's beliefs as a consequence of this labelling:

$$P(l|b,h) \propto P(b|l,h)^h, \tag{4}$$

where the normalizing term is obtained by summing over all possible labels l. Intuitively, the equation states that the learner expects the communicator to choose labels that tend to maximize the probability of the learner believing what the communicator believes in the helpful case, and minimize this probability in the unhelpful case. This is because when the communicator is being helpful, $h = 1$, they choose labels in such a way that tends to maximize $P(b|l,h)$, the probability of the belief they actually hold. When the informant is not being helpful $h = -1$, they choose labels in a way that tends to minimize the probability of the learner inferring the belief the communicator holds.[2] While Equation 4 does not directly tell us what kind of labeling distributions actually satisfy this relationship, it is possible to discover this using numerical methods like fixed point iteration (see Shafto & Goodman, 2008; Shafto *et al.*, under revision, for more detailed discussion).

In fact, although the underlying theory is complex, the behavior that results turns out to be quite simple. For instance, suppose that the informant can choose between four labels (A, B, C and D), and has the belief that the correct label is A. It turns out that, if the informant is trying to be helpful, then the probability that the informant will choose label A (that is, $P(l = A|b = A,h = 1)$) is high. The other three labels are all equally likely, and chosen with low probability.[3] On the

[2] Note that in this equation, we assume that learners have uniform prior expectations about the informant's possible beliefs.

[3] When the informant's knowledge and intent is known, these probabilities are 1 and 0, respectively.

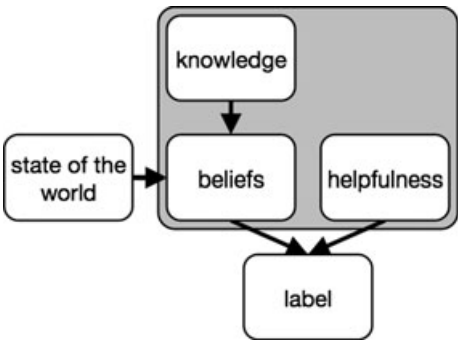

Figure 1 *Graphical representation of the sampling model. Boxes indicate variables, and lines indicate probabilistic dependence. The variables of knowledge k, helpfulness h, and beliefs b are all properties of the informant: the informant's beliefs depend on whether they are knowledgeable about the world, and their beliefs and and helpfulness jointly determine the label they choose.*

other hand, if the informant is being unhelpful, this situation reverses, with the probability of label *A* becoming low, and the probability of the other labels rising.[4]

Modeling results

Modeling 4 year-olds' behavior in trust tasks

We model the performance of 4 year-olds, contrasting the fits of the Knowledge & Intent model with the fits of a Knowledge-only model. In the experiments described above (Pasquini *et al.*, 2007; Corriveau & Harris, 2009a; Corriveau *et al.*, 2009a; Mascaro & Sperber, 2009), there are two types of question asked of children. Ask Questions query children about which informant they would rather ask for information. Endorse Questions occur in situations in which multiple informants provide labels for an unfamiliar object, and children are asked what they believe the object is called. The first two experiments (Pasquini *et al.*, 2007; Corriveau & Harris, 2009a) involved Ask Questions, which we model by evaluating which of the informants is more likely to provide the correct label. The second two experiments (Corriveau *et al.*, 2009a; Mascaro & Sperber, 2009) involved Endorse Questions, which we model by again asking it to choose between informants and assuming that the label is the one that the chosen informant previously generated. Full mathematical details can be found in Appendix A.

Our model, like children, makes inferences about the informant's knowledge k and helpful intent h; like children, inferences in the model are shaped by prior expectations about whether informants are likely to be helpful and/or knowledgeable. Our modeling framework allows us to explore the nature of the expectations that children have by allowing us to determine which parameters best fit their observed behavior. Each component, knowledge and intent, is captured by two parameters: balance and uniformity. Balance represents children's expectation about people on average, while uniformity captures the expected variability across individuals. For the Knowledge & Intent model, we infer whether children believe that people are generally knowledgeable and helpful, and whether people are uniform or variable in those tendencies.

These parameters have important implications for, among other things, children's ability to learn from evidence and consequently the changes in predictions across conditions. To the degree that children believe that people are uniformly knowledgeable or helpful, the model predicts that they will be relatively insensitive to evidence to the contrary. In contrast, if they expect people to vary in knowledge or helpfulness, the model predicts that they will be relatively sensitive to evidence, updating their beliefs about an individual given the information that they provide. Similarly, to the degree that children believe that people tend to be, but are not perfectly helpful, they will remain skeptical of even informants who provide truthful information, effectively reducing the variability in their behavior.

To model behavior based on knowledge alone, we must make some assumption about how knowledgeable people choose labels. Previous empirical research is not explicit about this issue, but a sensible approach is to model children as assuming that informants choose information helpfully. To capture this, we fix beliefs about helpfulness at $P(h) = 1$; that is, informants are assumed to be always helpful. The key question will be: is behavior adequately explained by inferences based on knowledge alone, or does the model with both knowledge and intent provide a significantly more accurate explanation of children's behavior?

The results reported here are based on a single set of best-fit parameters for the Pasquini *et al.* (2007), Corriveau and Harris (2009a), and Corriveau *et al.* (2009a) studies. For the Knowledge & Intent model, the best fitting parameters are .01 uniformity and .75 balance for knowledge, and .001 uniformity and .75 balance for helpfulness. The values of the balance parameters indicate that 4 year-olds assume that people are generally knowledgeable and helpful. The values of the uniformity parameters indicate that they assume that informants tend to be variable. Together the parameters indicate that informants are generally assumed to be knowledgeable and helpful, but there are informants who tend to be not knowledgeable and/or not helpful.

For the Knowledge-only model, the best fitting parameters are .01 uniformity and .05 balance. According to the Knowledge-only model, children assume that people as a group tend to be not particularly knowledgeable, but people do vary–individuals are either knowledgeable or not. These parameter values are surprising, and we discuss why these parameters best fit the data in greater detail below. Full details about the fitting procedure are reported in Appendix B.

[4] The exact probabilities are 0 and 1/3, assuming four labels.

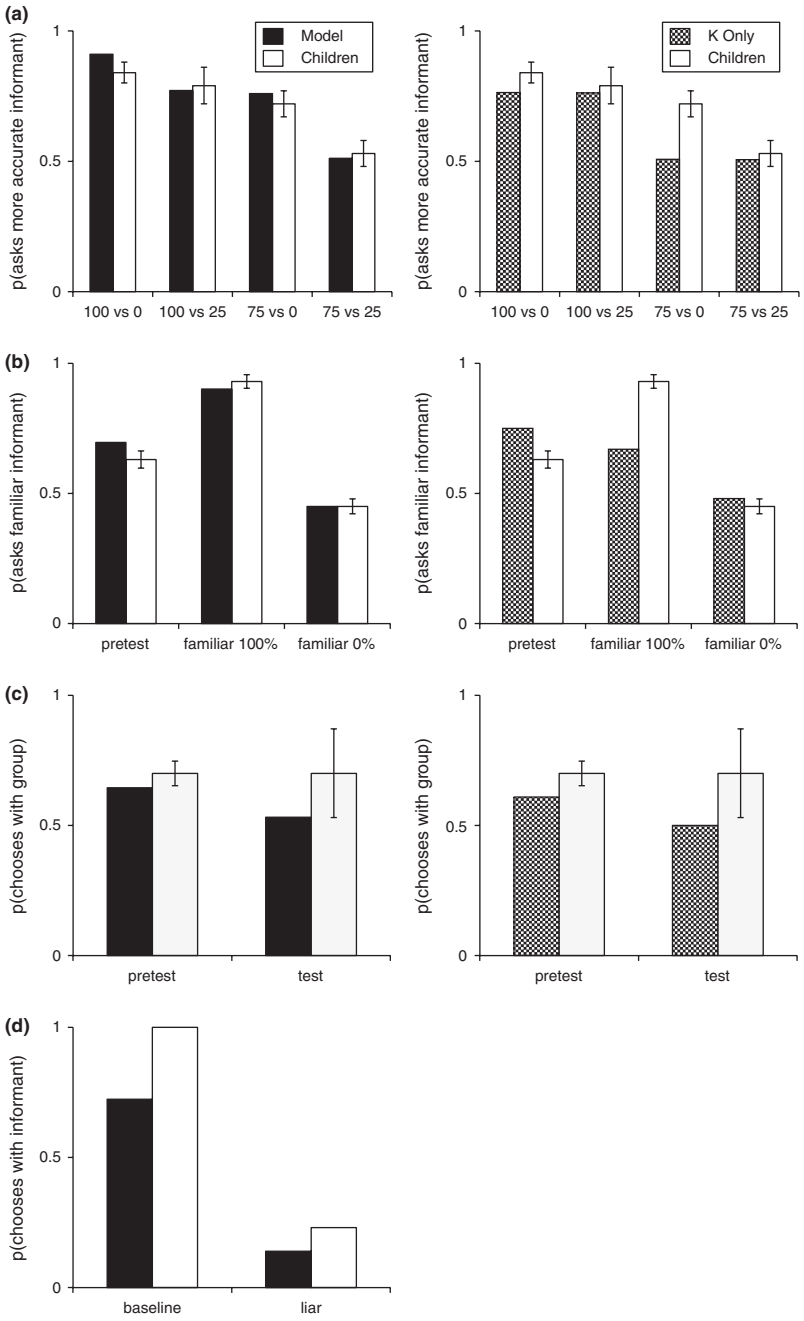

Figure 2 *Model predictions and observed results for 4 year-olds' choices in epistemic trust tasks. Black bars represent the predictions of the Knowledge & Intent model; gray bars represent predictions of the Knowledge-only model; white bars represent children's behavior.(a) In Pasquini et al. (2007), children were asked which of two informants they would trust to provide a new label. The informants differed in their accuracy on labels known to the children (shown on the x axes). Both children and the Knowledge & Intent model choose the more accurate informant, and the strength of the inference decreases when accuracy is probabilistic. The Knowledge-only model fails to capture the effects of variations in accuracy.(b) In Corriveau and Harris (2009), the children and both models initially prefer a familiar informant who has been known to be knowledgeable and helpful (*PRE-TEST*), and continue to prefer that informant if she ontinues to be accurate (*FAM 100%*). However, if she does not, they switch their preference to the novel informant (*FAM 0%*).(c) In Corriveau et al. (2009a) children are presented with a novel object that is labeled by groups of informants, a majority of whom (but not all) agree on the label. Both children and the Knowledge & Intent model use agreement among informants to infer that informants in the majority are more trustworthy. This is true both when asked about that label (*PRE-TEST*) as well as when deciding about a novel label provided either by the original dissenter or one of the original majority informants (*TEST*). Both models somewhat overestimate the difference between the questions; however, the Knowledge-only model performs considerably worse.(d) In Mascaro and Sperber (2009), after a baseline test, children were told that the informant was 'a big liar'. They, like the Knowledge & Intent model, use this information to choose consistently with the informant only if the informant is not a liar. The Knowledge-only model is omitted because it, by definition, cannot capture the effect of a knowledgeable but not helpful informant.*

For each model, we can assess the probability of the children's choices, and use model selection to decide whether children's behavior is best explained by the Knowledge-only model or the Knowledge & Intent model. For instance, for each model, we can assess the probability that 10 out of 15 children would choose a given informant.[5] Assessing the probability of the whole set of observed data (all choices in all of the studies), and comparing the probabilities of the observed data under each model (via a likelihood ratio test) provides a method by which we may determine whether children's behavior is best explained by inferences about knowledge and intent, or knowledge alone. A likelihood ratio test suggests that the Knowledge & Intent model provides significant additional explanatory value over the Knowledge-only model, $\chi^2(2) = 40.65, p < .001$.[6] In the following, we discuss the model fits for each study in detail, exploring how the models differ in their fits to children's behavior.

Formerly accurate/inaccurate informants

Several papers have focused on how children react to informants they have observed either correctly or incorrectly label familiar objects (Koenig & Harris, 2005a; Pasquini *et al.*, 2007; Corriveau *et al.*, 2009b). In these studies, the informants first label objects whose labels are known to the child (e.g. BALL or SHOE). The informants then give novel objects an unfamiliar label. In the most common case, one informant has labeled all of the familiar objects correctly, while the other has labeled them all incorrectly. The critical question is whether children can infer who to trust based on the evidence. We focus on the results of Pasquini *et al.* (2007), which included the contrast between perfectly accurate (knowledgeable) and perfectly inaccurate (not knowledgeable) informants but also explored situations involving partially accurate informants, which provides a much richer data set to test the models against. To do so, we include their results from Experiments 1 and 2 (for the 75% versus 0% condition we used the data from the condition with more participants, Experiment 2).

The Knowledge & Intent model predicts a strong preference for the accurate informant when one is 100% accurate and the other is 0% accurate. However, when one is 75% accurate and the other is 25% accurate, the preference for the more accurate informant is weak-to-nonexistent. For the two interim conditions (100% versus 25% and 75% versus 0%), the more accurate informant is preferred by the model, but there is little difference in preference between these two conditions. Specifically, in the 75% versus 0% condition, the informant getting one incorrect is clearly not unhelpful and should be trusted over the always inaccurate informant. These results closely match the qualitative and quantitative trends in children's behavior, as shown in Figure 2a.

In contrast, the Knowledge-only model fails to capture the gradual drop in trust. The parameter values embody the assumption that people tend to be either knowledgeable or not, though most are not knowledgeable (low uniformity and balance parameters). While these values seem counter-intuitive, they reflect a fundamental contradiction in the data from the perspective of a Knowledge-only model. If the model assumes that people tend to be either knowledgeable or not (as indicated by a low value of the uniformity parameter), then wrong answers become more diagnostic, because knowledgeable people will be 100% correct (because they are always helpful). This leads the model to show a strong preference when there is a 100% correct informant, and relative indifference otherwise. Alternatively, the model could assume that people are uniform in their degree of knowledge (high uniformity parameter), and people's knowledge may not be perfect. In this case, the model would have a very hard time explaining any differences at all because informants' knowledge would be uniform. A third possibility is to assume that degree of knowledgeability varies both within and across individuals (uniformity parameter near 1). Again, however, wrong answers are more diagnostic of lack of knowledge than correct answers are of knowledge; the model predicts that 100% versus 25% would be similar to 100% versus 0% and both would be greater than 75% versus 0% and 75% versus 25% (and there may be an above-chance preference for the 75% correct informant). Thus, the best fit to children's behavior is the low value of the uniformity parameter; it predicts strong preferences in the conditions where there is an informant who is 100% correct and indifference in the 75% versus 25% condition, but is far off in the 75% versus 0% condition (the exact value of the balance parameter tends not to make much difference in this case).

Familiar informants

In Corriveau and Harris (2009a), children were asked to choose between a new informant and a familiar, previously trustworthy informant (their preschool teacher). As a measure of who the child showed a prior preference for, the children were given a pre-test in which their teacher and the novel informant both labeled novel objects and children were asked who they would prefer to ask for the label (the pretest values for the two conditions were averaged).[7] Children in one condition (FAM 100%) then

[5] Because in some of the papers each child made more than one decision but only aggregated probabilities were presented, we assume that each observation is independent to allow calculation of the relevant frequencies.

[6] This result is based on four labels. The differences between the models hold over a range of values (see Appendix B).

[7] The study also used novel functions, but we do not model those data here.

saw the familiar informant label four familiar objects correctly, but the new informant labeled them incorrectly. In the other condition (FAM 0%), the new informant labeled them correctly and the familiar one did not. Finally, the child was presented with a novel object, and children were asked who they would ask to label the object.

Modeling these results requires incorporating children's extensive past experience (presumably positive) with their teacher into the model. To do so, we provided the model with 20 demonstrations by the informant, 19 of which were knowledgeable and 19 of which were helpful (see Appendix B for mathematical details). Before this update, prior biases about the novel informant were the same as in all other studies. The model qualitatively captures all of the empirical findings, as shown in Figure 2b. Because the familiar informant is believed to be more helpful and knowledgeable than the new one, both children and the model prefer the familiar informant during pre-testing. Both children and the model are also able to use accuracy on the known labels to make inferences about both informants; as a result, both favor the familiar informant when the familiar informant is accurate but prefer the new informant if the familiar informant is not accurate.

In contrast, the Knowledge-only model fails to capture the increase in trust that results from the familiar informant performing well. Because past experience has been favorable, the model already strongly infers that this person must be knowledgeable, so a few added data points do not change the predictions much, leading to a sizable deviation from children's behavior.

Groups of informants

So far we have seen studies in which children observed informants labeling known objects. They and our model were able to use this information to make inferences about the informants. However, in other situations children may not know the correct label: how are they to decide which informants to believe? Corriveau *et al.* (2009a) investigated whether children could use information about the degree of agreement between informants when determining who to trust. For example, given a novel object and a group of four informants, if all except a single dissenter agree on which object corresponds to the label, whose information will the child trust?

The Knowledge & Intent model infers that the answer chosen by the majority is more likely to be correct. This is because (as long as there is no collusion) the probability that a group of non-knowledgeable or non-helpful informants would randomly converge on a single answer is low. As shown in Figure 2c (PRE-TEST), this is the same inference that children make. One can also explore the robustness of the inferences made about the informants by having the dissenter and one of the majority informants each provide a different label for a new object. In this situation both children and the Knowledge & Intent model have a slight preference for the label provided by the informant from the majority, though the model underestimates the strength of the effect (Figure 2c, TEST).

The Knowledge-only model also captures the basic effect, for the same reason. The chances of three not-knowledgeable informants agreeing is very low, and as a consequence the model predicts that children should trust the group and informants from the group in future encounters.

At this point, it is worth revisiting why the Knowledge-only model chooses the parameters it does. Recall that the best fitting parameters were .01 uniformity and .05 balance, suggesting that the best explanation of children's behavior according to the Knowledge-only model was that 4 year-olds do not believe that people tend to be knowledgeable. A problem arises in that the Knowledge-only model has a difficult time explaining two aspects of children's behavior simultaneously. Children appear to take consensus as a relatively weak indicator for trust, suggesting that people are generally not knowledgeable. For the familiar informants experiment, children appear to be highly sensitive to negative evidence; a few incorrect answers from a familiar and trusted informant leads to a reversal of trust. Both of these point to a strong expectation that people generally produce incorrect answers. To explain the remaining results, there must be some possibility that people are knowledgeable. Together, these results are only consistent with generally not-knowledgeable informants, with high variability across individuals.

In contrast, the Knowledge & Intent model infers that people tend to be both knowledgeable and helpful, but individuals do vary. Joint inferences about knowledgeability and helpfulness allow for intuitive explanations for these otherwise surprising events. For instance, how should one explain an informant labeling four out of four familiar objects incorrectly? It could be that they are not knowledgeable; however, it seems reasonable and plausible that they are deceiving. This possibility would capture children's graduated reaction to evidence about reliability on familiar objects, and their ability to use information about both correct and incorrect responses by familiar informants. Similarly, because there is ambiguity about knowledge and helpfulness, the Knowledge & Intent model naturally captures children's not complete trust in the group's response. In sum, children's behavior is best explained by inferences about both knowledge and intent.

Deceptive informants

How do children reason if they know an informant is deceptive? Mascaro and Sperber (2009) explored this by presenting children with a knowledgeable informant whom they were told was deceptive. A piece of candy was secretly placed under one of two boxes. The informant (a

puppet) looked under both boxes in view of the child, thus alerting the child to the fact that the informant was knowledgeable. The experimenter warned the child that the informant 'always tells lies', after which the puppet indicated verbally which box the candy was hidden under and the child was asked to guess which box had the candy. Note that this problem is similar to the labeling problems above–the puppet is conveying information about the state of the world. The critical question was whether children knew to choose the opposite box. Indeed, 4 year-olds chose the opposite box approximately 77% of the time, in contrast with their responses at baseline, after watching the puppet look under a different set of boxes, but before being told about the puppet's lying ways, where they looked under the cup he pointed to 100% of the time.

We model baseline performance by generating predictions about which of the two boxes to choose using the same helpfulness parameters as the other experiments, but with the knowledgeability parameter set to be high[8] to capture the fact that the puppet, having looked under the cups, knew where the candy was. The model predicts that children should tend to trust the informant, though the strength of the model's prediction is weaker than children's. We model performance after being told that the puppet looked under the cups and always lies by retaining the high prior probability of knowledgeability but changing the prior probability to favor deceptiveness.[9] Figure 2d shows that the model captures the reversal of choice after learning about the informant's deceptive ways.

The Knowledge-only model predictions are omitted for these data because, by definition, it cannot capture how information about the informant's intent affects inference.

Developmental changes between 3 and 4 years old

These results suggest that the Knowledge & Intent model provides the best explanation of 4 year-olds' behavior in epistemic trust tasks. However, there are developmental changes from age 3 to 4. In some instances, 3 year-olds show the same qualitative behavior as 4 year-olds (see Figure 3c). In others, 3 year-olds show qualitative differences (Figure 3a, b, and d). This raises a question: what causes the changes in behavior between the ages of 3 and 4?

One possibility is that at age 3, children are relying only on inferences about knowledge. Indeed, the experiment in which 3 year-olds are most different from 4 year-olds (Mascaro & Sperber, 2009) relies critically on using information about an informant's intent; 3 year-olds choose to trust an informant that they are told is a liar, while 4 year-olds do not. Comparing the fits of the Knowledge & Intent model and the Knowledge-only model to the other three sets of results suggests that the Knowledge & Intent model does not capture significant additional variance, $\chi^2(2) = 0.03$, $p > .5$, consistent with the explanation that 3 year-olds do not use inferences about intent to decide whom to trust.[10]

Perhaps most tellingly, we can compare the best fitting parameter values for each model. For the Knowledge-only model, the best fitting parameters are 2.00 for uniformity and 0.05 for balance. For the Knowledge & Intent model, the best fitting parameters are 2.00 for knowledge uniformity, 0.05 for knowledge balance, 2.00 for helpfulness uniformity, and 0.95 for helpfulness balance. Note that the helpfulness parameters correspond to a relatively strong assumption that people are uniformly knowledgeable and helpful (as in the Knowledge-only model), and the knowledge parameters are identical. As a consequence, the models show similar fits to the data (see Figure 3).

General discussion

We have presented a model of epistemic trust as inference about the knowledge and intent of informants. Previous research focused on inferences about informants' knowledge to explain 3 and 4 year-olds' behavior on epistemic trust tasks. A parallel line of research using different methods provided evidence for developmental changes in whether children use information about informants' intent to deceive to guide reasoning. We have shown that inferences about knowledge alone do not account for the empirical results in epistemic trust tasks; a model that simultaneously makes inferences about knowledge, intent, and the state of the world provides a better fit to 4 year-olds' behavior. Moreover, the model suggests that developmental changes in behavior on these tasks stem from changing assumptions about informants' helpfulness. Taken together, these results provide evidence that epistemic trust depends on inferences about both informants' knowledge and their intent, and that changes in behavior on epistemic trust tasks are primarily due to changes in beliefs about intent.

Research on epistemic trust has manipulated children's beliefs about informants' knowledge primarily by varying the data that the informants provide (Corriveau & Harris, 2009a; Corriveau *et al.*, 2009a; Pasquini *et al.*, 2007). We have highlighted the fact that this information is ambiguous; informants may be wrong because they are not knowledge, unhelpful, or

[8] That is, we set $\beta_k = 0.9$.

[9] This corresponds to $\beta_k = 0.9$ and $\beta_h = 0.1$. The uniformity of both of these beliefs was set to be high (10), which is appropriate because the children watched the puppet look under the cups and were explicitly told that the informant was a liar. As detailed in Appendix B, the model fit is robust across a wide range of parameters that capture these intuitions.

[10] This result is based on four labels. The non-significant difference holds over a range of values (see Appendix B).

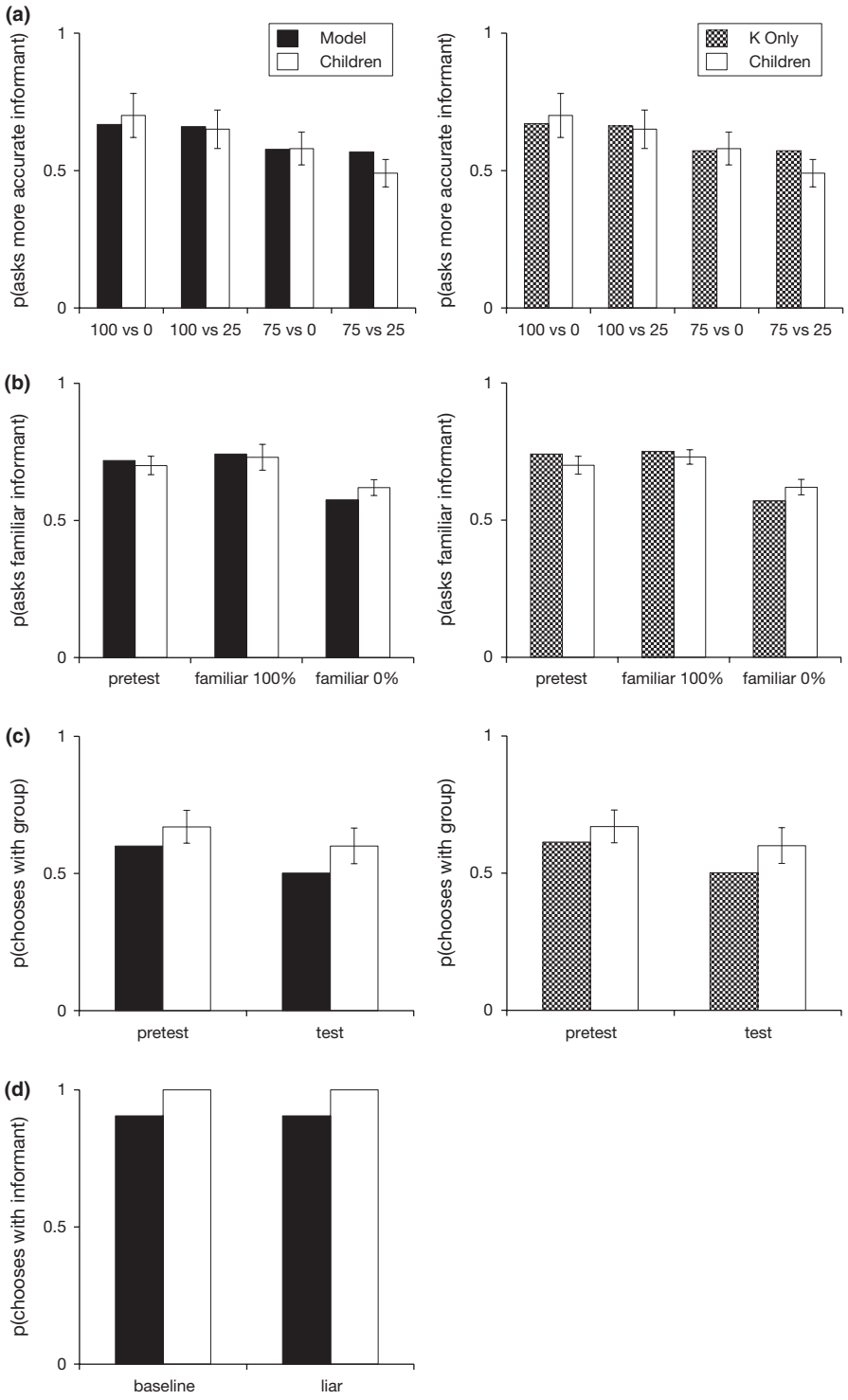

Figure 3 *Model predictions and observed results for the performance of three year-olds in the same epistemic trust tasks as in Figure 2. Across most tasks, three year-olds' performance is similar to that of four year-olds, with the exception of inferences about deception (d). In that experiment, three year-olds reliably choose the label provided by the informant regardless of whether that informant is called a liar or not. The model explains this developmental shift as a difference in prior expectations about helpfulness. If three year-olds have a much stronger expectation that people are helpful, then the information about the informant's intent would have less effect on their inferences, leading to qualitatively different behavior.*

both. Our results show that these two components are necessary to account for children's behavior on these tasks and there are developmental changes in these abilities. We cannot say whether these developmental changes are specific to epistemic trust tasks or whether they reflect a change in children's competencies. Answers to this question will require new methods for isolating and manipulating attributions of knowledge

and helpfulness. Interestingly, our analysis suggests that some tasks, such as groups of informants (Corriveau *et al.*, 2009a), mainly elicit attributions of knowledgeability. An important direction for future research is to continue to explore methods that separate attributions of knowledge from attributions of helpfulness and how these abilities change over development (see Mascaro & Sperber, 2009; Vanderbilt, Liu & Heyman, 2011).

The modeling approach presented here is related to recent research investigating the role of pedagogical inferences in learning (Bonawitz, Shafto, Gweon, Goodman, Spelke & Schulz, 2011; Buchbaum, Griffths, Gopnik & Shafto, 2011). These experiments manipulated children's beliefs about helpful demonstrators' knowledge and investigate the inferences that children draw from these demonstrations. The results show that preschool-aged children draw qualitatively different inferences from the same data, depending on whether the informant was helpful and knowledgeable or not knowledgeable. Furthermore, children's inferences in these experiments are well-predicted by a model of pedagogical reasoning that uses the fact that an informant is knowledge and has helpful intent to facilitate learning (see Shafto & Goodman, 2008). The model proposed here is a generalization of that approach, where knowledge and intent are inferred rather than assumed; therefore, these previous demonstrations provide additional support for the current model. This also suggests that an important avenue for future research is to investigate how the manipulations used in epistemic trust tasks affect learning in ways that go beyond simply endorsing information; how do incorrect labeling, dissent, and familiarity affect inferential learning?

Our model is a computational-level account (Anderson, 1990; Marr, 1982) that provides a formal, rational analysis of the problem of epistemic trust. Our goal was to describe how a learner might combine inferences about an informant (specifically, their knowledge and intent) with data about the state of the world (in this case, labels for objects) to simultaneously learn the true state of the world and which informants might be trusted. We make no claims about the kinds of mechanisms that may implement these computations in the brain. Nevertheless, our model provides insight into the developmental processes that may underlie emerging competence in epistemic trust, and suggests that the changes in behavior observed between 3 and 4 years of age may result from changes in children's ability to reason about intent. An interesting question for future empirical research is whether different experiences or more experience is related to these changes.

Our account of deception, based on the experiments by Mascaro and Sperber (2009), considered only an informant who 'always lies'. In the real world, of course, few people always lie–even if the intent is always to mislead, informants may sometimes tell the truth in order to deceive. Extensions of our framework to capture richer notions of deception are possible by allowing informants to modify their behavior based on their inferences about the learner's inference, and we leave these extensions to future work.

Similarly, we have focused on the basic phenomena related to epistemic trust. Recent work has investigated how other factors, including effects of perceptual appearance of the object (Corriveau & Harris, 2010) and the speaker's accent (Kinzler, Corriveau, & Harris, in press) interact with epistemic trust. Extending the model to incorporate these other influences is an important direction for future work.

In sum, this research provides a novel formal account of epistemic trust as well as an exploration of the changing nature of epistemic trust over development. Together with recent empirical and modeling results, our account suggests that social understanding is a crucial component of children's learning and development. Understanding the richness of inferences about others' knowledge and intent is therefore necessary to understanding the power of human learning.

Acknowledgements

This work was partially funded by University of Louisville IRIG-URG grant to P.S., a subaward from the James S. McDonnell Foundation to P.S., and ARC grant DP0773794 to D.J.N.

References

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Bonawitz, E.B., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Teaching limits children's spontaneous exploration and discovery. *Cognition*, **120**, 322–330.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistimology*. Oxford: Clarendon Press.

Buchbaum, D., Griffths, T., Gopnik, A., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, **120**, 331–340.

Clement, F., Koenig, M., & Harris, P.L. (2004). The ontogenesis of trust. *Mind and Language*, **19**, 360–379.

Corriveau, K.H., Fusaro, M., & Harris, P.L. (2009a). Going with the flow: preschoolers prefer non-dissenters as informants. *Psychological Science*, **20**, 372–377.

Corriveau, K.H., & Harris, P.L. (2009a). Choosing your informant: weighing familiarity and past accuracy. *Developmental Science*, **12**, 426–437.

Corriveau, K.H., & Harris, P.L. (2009b). Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *Developmental Science*, **12**, 188–193.

Corriveau, K. H., & Harris, P. L. (2010). Preschoolers (sometimes) defer to the majority when making simple perceptual judgments. *Developmental Psychology*, **26**, 437–445.

Corriveau, K.H., Meints, K., & Harris, P.L. (2009b). Early tracking of informant accuracy and inaccuracy by young children. *British Journal of Developmental Psychology*, **27**, 331–342.

Guan, Y., Fleibner, R., Joyce, P., & Krone, S. (2006). Markov chain Monte Carlo in small worlds. *Journal of Statistics and Computing*, **16**, 193–202.

Harris, P.L. (2007). Trust. *Developmental Science*, **10**, 135–138.

Harris, P.L., & Corriveau, K.H. (in press). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B*.

Jaswal, V.K., Croft, A.C., Setia, A.R., & Cole, C.A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, **21**, 1541–1547.

Jaswal, V.K., & Neely, L.A. (2006). Adults don't always know best: preschoolers use past reliability over age when learning new words. *Psychological Science*, **17**, 757–758.

Kinzler, K.D., Corriveau, K.H., & Harris, P.L. (in press). Preschoolers' use of accent when deciding which informant to trust. *Developmental Science*.

Koenig, M.A., & Harris, P.L. (2005a). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, **76**, 1261–1277.

Koenig, M.A., & Harris, P.L. (2005b). The role of social cognition in early trust. *Trends in Cognitive Sciences*, **9**, 457–459.

Marr, D. (1982). *Vision*. New York: W.H. Freeman.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, **112**, 367–380.

Pasquini, E.S., Corriveau, K.H., Koenig, M.A., & Harris, P.L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, **43**, 1216–1226.

Shafto, P., & Goodman, N.D. (2008). Teaching games: statistical sampling assumptions for pedagogical situations. In Proceedings of the 30th annual conference of the Cognitive Science Society.

Shafto, P., Goodman, N.D., & Griffths, T. L. (under revision). Rational reasoning in pedagogical contexts. Manuscript under review.

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

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2010). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems*, **22**, 1874–1882.

Vanderbilt, K.E., Liu, D., & Heyman, G.D. (2011). The development of distrust. *Child Development*, **82**, 1372–1380.

Received: 3 February 2011
Accepted: 22 December 2011

Appendix A: Knowledge & Intent model specifications and inference algorithm

In the model described in the text, the learner has three key prior beliefs that need to be specified. First, we need to specify the learner's prior bias to believe that the informant will be helpful. Similarly, we need to describe the bias to belief that the informant is knowledgeable. Finally, we need to specify, the learner's prior knowledge about the true state of the world.

Consider the learner's beliefs about helpfulness. We want to be able to model these beliefs at three different levels: general expectations about people, the specific informant's tendencies, and whether an informant is knowledgeable and/or helpful on a particular trial. To capture these distinctions in a probabilistic generative model, we begin by assuming that there is some probability $\theta_h = P(h = 1)$ that describes the chance that the informant will be helpful on any specific trial. In statistical notation, this is written

$$h \sim Bernoulli(\theta_h). \tag{5}$$

Thus, h describes whether the informant is being helpful on this particular trial, whereas θ_h describes the overall tendencies of this particular informant. To capture the idea that the learner has some more general beliefs about people, we assume that there is a Beta distribution over θ_h, which is parameterised by β_h, the learner's bias to believe that people are usually helpful which we call *balance*, and γ_h, a parameter that describes whether people are uniformly helpful or whether different people differ in helpfulness:

$$\theta_h \sim Beta(\gamma_h\beta_h, \gamma_h(1-\beta_h)). \tag{6}$$

Following the same logic, we can specify the learner's beliefs about the knowledgeability of informants in much the same way:

$$k \sim Bernoulli(\theta_k) \tag{7}$$

$$\theta_k \sim Beta(\gamma_k\beta_k, \gamma_k(1-\beta_k)) \tag{8}$$

and unless otherwise specified in the text, both of the balance parameters β_h and β_k and the uniformity parameters γ_h and γ_k were treated as free parameters that we fit to the data for the Knowledge & Intent model (details about parameter fitting can be found in Appendix B).

To make inferences about who the learner should ask for information, we assume that informants are chosen with probability proportional to the chance that they will actually choose the correct label. Accordingly, we need to calculate, for all informants, the probability that the informant will provide the correct label (i.e. $l = s$), conditioned on the learner's previous experience with that informant (denoted E), and also taking the learner's generic prior biases about people into account. This is given by:

$$P(l = s|E, \beta, \gamma) = \sum_s P(s) \int P(l = s|s, \theta)P(\theta|E, \beta, \gamma)d\theta \tag{9}$$

where $\theta = (\theta_h, \theta_k)$ refers to both helpfulness and knowledgeability of the informant, $\beta = (\beta_h, \beta_k)$ refers

to the learner's biases about helpfulness and knowledgeability, and $\gamma = (\gamma_h, \gamma_k)$ refers to the uniformity of these two biases. In this equation, the outer summation is taken over all possible states of the world (i.e. all possibilities as to the identity of the true label), and the integration is taken over all possible values of θ_h and θ_k (i.e. from 0 to 1 for both variables).

While the $P(s)$ term in this expression is very simple, and the summation over all possible values of s is similarly straightforward (we assume for simplicity that there are four possible labels), the integration in Equation 9 is non-trivial, and is certainly analytically intractable. The difficulty of this inference becomes clear when it is recognized that $P(\theta|E,\beta,\gamma)$ involves calculating the posterior distribution over possible helpfulness and knowledgeability rates in light of all previous experiences. As a consequence, we use Monte Carlo methods (in this case, rejection sampling) to numerically approximate the probability $P(l = s|E,\beta,\gamma)$ that a particular informant will give the correct label.

The description above assumes that the learner's goal is to decide which informant to request information from. However, we can capture 'endorse' questions simply by conditioning on a particular state of the world; the change in prediction is relatively minor.

Appendix B: Parameter fitting and model evaluation

To find the best fitting parameters for the 4 year-olds' data for the Knowledge & Intent model, we performed a grid search over the values of γ_h, γ_k, β_h, and β_k. The uniformity parameters, γ_h and γ_k, may take on values between 0 and infinity. When γ_h is near zero, the expectation is that people tend to be very helpful or very deceptive, with the average indicated by the β_h; whereas, when γ_h is large, the expectation is that each individual is well characterized by the β_h parameter. Grid search for the best values of γ_h and γ_k was performed on a roughly logarithmic scale, over the values [0.001, 0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0]. The balance parameters, β_h and β_k, may take on values between 0 and 1. When they are near 1, the expectation is that people as a group tend to be helpful (or knowledgeable). When they are near 0, the expectation is that people tend to be deceptive (or not knowledeable). Grid search was performed over the values [0.95, 0.75, 0.5, 0.25, 0.05].

Because Corriveau and Harris used familiar informants who were known to be knowledgeable and helpful in the past, for this data set these parameters were updated to reflect the prior experience. To capture this, we imagine 20 additional observations in which the familiar informant was helpful 19 times and knowledgeable 19 times. Note that these numbers are chosen to represent simply having more experience, and the results are robust across a range of parameters capturing this basic intuition (5–200 observations with proportions of helpfulness and knowledgeablity greater than .5). Due to the properties of the Beta distribution, we can update beliefs about this person using simple algebra. For instance, the updated beliefs about this individual β'_h is simply $\frac{\beta_h\gamma_h + 19}{\gamma_h + 20}$, and the updated γ'_h is γ_h + 20.

At each set of parameter values, the sum of squared distance between the observed and predicted value was computed for the Pasquini *et al*. (2007), Corriveau and Harris (2009a), and Corriveau *et al*. (2009a) data. The results based on the parameter set with with the smallest mean squared error for the 4 year-olds' data, $\gamma_k = 0.01$, $\beta_k = 0.75$, $\gamma_h = 0.001$, and $\beta_h = 0.75$, are shown in Figure 2. The results based on the parameter set with the smallest mean squared error for the 3 year-olds' data, $\gamma_k = 2.00$, $\beta_k = 0.05$, $\gamma_h = 2.00$, and $\beta_h = 0.95$, are shown in Figure 3.

For the Knowledge-only model, the procedure was the same with the exceptions that the γ_h and β_h were set to 1 and 1000000 to capture the assumption that people are helpful. The results based on the parameter set with with the smallest mean squared error for the 4 year-olds data, $\gamma_k = 0.01$ and $\beta_k = 0.05$, are shown in Figure 2. The results based on the parameter set with the smallest mean squared error for the 3 year-olds' data, $\gamma_k = 2.00$ and $\beta_k = 0.05$, are shown in Figure 3.

The Mascaro and Sperber (2009) data were fit separately. In Mascaro and Sperber children were told that the informant always tells lies, and children observed the informant looking under the cups. To capture these two manipulations of children's beliefs about this informant in the Knowledge & Intent model, we set $\beta'_h = .1$, $\beta'_k = .9$, and the uniformity to be high for both, 5. In both cases, the results were robust across a range of values for the uniformity parameters (2–100) and values of the balance parameters consistent with the notion that the informant is knowledgeable and unhelpful (.6–99 for knowledge and .01–4 for helpfulness).

Finally, to ensure that the results reported in the paper generalized to different numbers of possible labels, we ran simulations in which we varied n from 4 to 128. As the value of n increased, the rejection sampling method described above performed less well. To conduct these simulations we implemented an MCMC algorithm to identify the best fitting model at each value of n (the specific approach was based on Guan, Fleibner, Joyce, & Krone, 2006). The results confirm the main result of the paper: 4 year-olds' behavior is significantly better explained by the Knowledge & Intent model, while 3 year-olds' is well explained by knowledge alone.