Patrick Shafto, Baxter Eaves, Daniel J. Navarro and Amy Perfors

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

Developmental Science, 2012; 15(3):436-447

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

Patrick Shafto & Baxter Eaves
University of Louisville

Daniel J. Navarro & Amy Perfors
University of Adelaide

Abstract

A core assumption of many theories of development is that children can learn indirectly from other people. However, unlike direct experience, indirect experience (or testimony) is not constrained to be veridical. 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 three and four years of age as a result of changing default assumptions about the helpfulness of other people.

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 indirect experience is potentially very informative, it also poses a problem: while direct experience always provides veridical information, indirect experience may not. 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, 2009; Corriveau, Fusaro, & Harris,

This work was partially funded by University of Louisville IRIG-URG grant to P.S. and AFOSR grant FIXME: airforce grant to A.P., D.N., and P.S.

2009; Corriveau & Harris, 2009a; M. A. Koenig & Harris, 2005; Clement, Koenig, & Harris, 2004; M. Koenig & Harris, 2005; Harris & Corriveau, in press; Corriveau & Harris, 2009b; Harris, 2007; Jaswal, Croft, Setia, & Cole, 2010; Jaswal & Neely, 2006).

Even preschool children can make sensible inferences about other people's knowledge. Koenig & Harris (2005) 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 four years of age, children are quite sophisticated in their ability to monitor others' knowledge. 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 1). 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, 2009). In addition, children use consensus as an indicator of knowledgeability, trusting informants who label objects in the same way as the majority of others (Corriveau et al., 2009). Together, these studies provide strong evidence that children monitor others' knowledge and use this information to infer who to trust.

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. In an experiment by Mascaro & Sperber (2009), children were presented with a situation in which there were two cups, under one of which was hidden a candy. An informant then looked under both cups, 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 cup by pointing. Their results showed that by four-years old, children used information about the informant's intent, showing a significant preference for the cup that the informant did not point to. Epistemic trust, even for preschoolers, is more than just monitoring knowledge.

We propose a theoretical framework for epistemic trust that 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 four-year-olds' behavior. Our model also suggests that developmental differences between the ages of three and four can be explained as a result of changing prior expectations about others' intent.

We proceed by introducing the model, which formalizes the relationship between evidence, epistemic trust, and learning. Next, we demonstrate that our model of trust captures four-year-olds' abilities but simpler models based on knowledge or intent alone do not. We also use the model to explain developmental differences between age three and four, and then 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

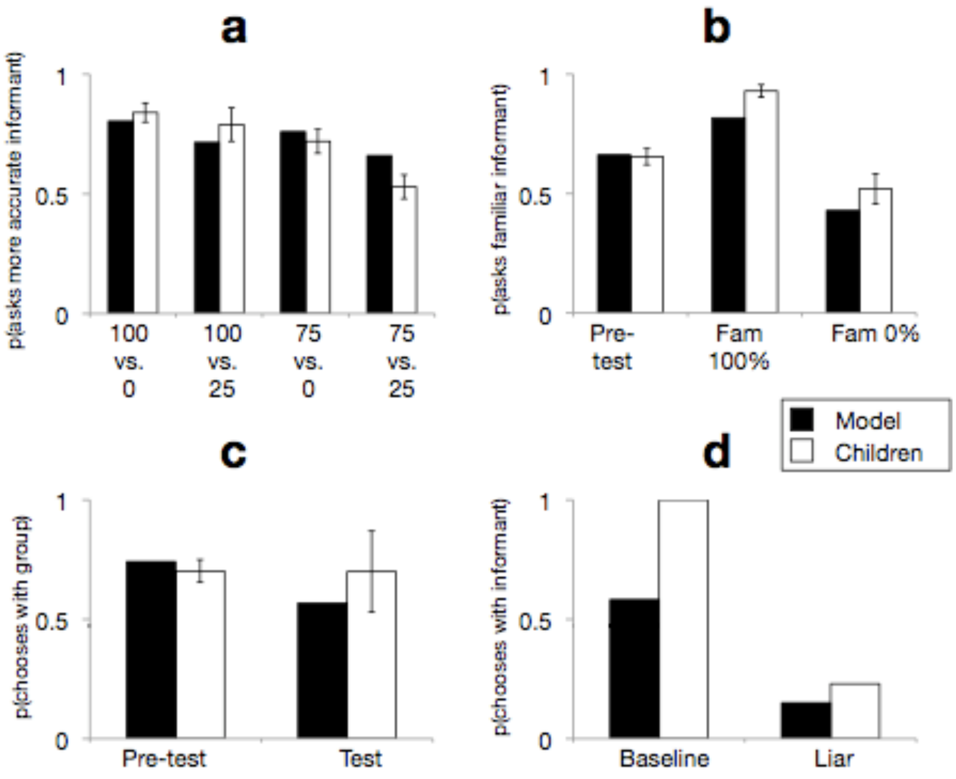

Figure 1. Model predictions and observed results for four-year-olds choices in epistemic trust tasks. (a) In Pasquini et al. (2007), children were asked which of two informants they would trust to provide a new label. The informants differed in their accuracy on labels known to the children (shown on the x axis). Both children and the model choose the more accurate informant, and the strength of the inference decreases when accuracy is probabilistic. (b) In Corriveau & Harris (2009), both children and the model initially prefer a familiar informant who has been known to be knowledgeable and helpful (PRE-TEST), and continue to prefer that informant if she continues 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 (2009) 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 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). (d) In Mascaro & Sperber (2009), after a baseline test, children were told that the informant was "a big liar". They, like the model, use this information to choose consistently with the informant only if the informant is not a liar.

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 all involve learning the correspondence between objects and their labels, 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, Griffiths, & 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 would explain the data (the likelihood). In this case, the beliefs we seek to model include childrens' 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, and we shall see in a later section that developmental differences between three and four years of age may depend on differences in these priors. 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, given the labels, can therefore reason backwards about all three of these things, given certain (sampling) assumptions about how helpfulness 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 $P(s, k, h \,|\, \ell)$, which according to Bayes' rule is given by:

$$P(s, k, h \,|\, \ell) \propto P(\ell \,|\, s, k, h) P(s, k, h), \tag{1}$$

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 ($h = -1$). 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 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 is not known, the learner 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 $P(s, k, h \,|\, \ell)$ is supplied by the likelihood function $P(\ell \,|\, s, k, h)$. 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 2 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, $P(b \mid k, s)$. 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 $P(\ell \mid b, h)$. Because the learner cannot directly observe the beliefs b of any informant, he 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:

$$P(\ell \mid s, k, h) = \sum_{b} P(\ell \mid b, h) P(b \mid k, s). \tag{2}$$

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:

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

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 $P(\ell \mid b, h)$, 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 $P(\ell \mid b, h)$ that satisfies the following "communicative sampling" relationship (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, under review). 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:

$$P(\ell \mid b, h) \propto P(b \mid \ell, h)^{h}, \tag{4}$$

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

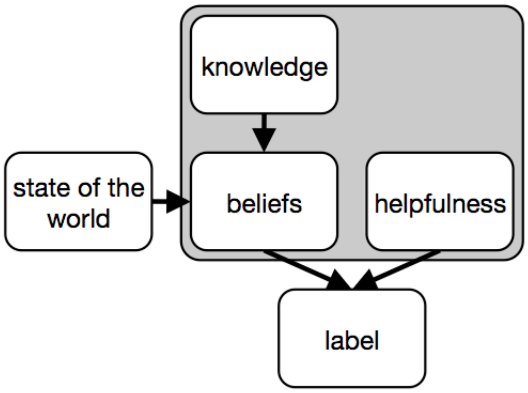

Figure 2. Graphical representation of the sampling model. Boxes indicate variables, and lines indicate probabilistic dependence. The variables of knowledge k, helpfulness h, and beliefs b are all properties of the informant: the informant's beliefs depend on whether they are knowledgeable about the world, and their beliefs and and helpfulness jointly determine the label they choose.

maximize $P(b|l,h)$, the probability of the belief they actually hold. When the informant is not being helpful $h = 0$, they choose labels in a way that tends to minimise the probability of the learner inferring the belief the communicator holds. [1] While Equation 4 does not directly tell us what kind of labelling 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 review, 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. On the 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.

Modeling results

We model the performance of four-year-olds in the experiments described above (Pasquini et al., 2007; Corriveau & Harris, 2009; Corriveau et al., 2009; Mascaro & Sperber, 2009). In these experiments, there are two types of questions 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, 2009) involved Ask Questions, which we model by evaluating which of the informants is more likely to provide the correct

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

label. The second two experiments (Corriveau et al., 2009; 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 biases about whether informants are likely to be helpful and/or knowledgeable. Our modeling framework allows us to explore the nature of the biases (if any) that children have, by allowing us to determine which biases best explain their observed behavior. Biases in the model correspond to parameters that describe the learner's beliefs about the informant's likely knowledge k and helpful intent h. We systematically vary these parameters and identify the values that best fit children's behavior. The results reported here are based on a single set of best-fit parameters for all of the studies, in which the prior bias over both helpfulness and knowledgeability is 0.6.[2] This can be interpreted as a moderate bias to think that informants will be both helpful and knowledgeable. Full details about the fitting procedure and cross-validation tests are reported in Appendix B.

Formerly accurate/inaccurate informants

Several papers have focused on how children react to informants they have observed either correctly or incorrectly label familiar objects (M. A. Koenig & Harris, 2005; Pasquini et al., 2007; Corriveau, Meints, & Harris, 2009). 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 our model against.

The 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 somewhat preferred by the model, and there is little difference in preference between these two conditions. These results closely match the qualitative and quantitative trends in children's behavior, as shown in Figure 1a.

Familiar informants

In Corriveau & Harris (2009), 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 naturally preferred, the children were given a pre-test in which their teacher and the novel informant both labeled novel objects and children were ask who

[2] In terms of the model details in Appendix A, this corresponds to the bias parameters β_h and β_k, with the strength of that bias represented by γ_h and γ_k (set to 1 as a default throughout).

they would prefer to ask for the label. Children in one condition (Fam 100%) then saw the familiar informant label three 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 with their teacher into the model; this is reflected in a strong prior belief that the familiar informant is both knowledgeable and helpful.[3] 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 1b. 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 were 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.

Groups of informants

So far we have seen studies in which children were able to see the informants label objects for which the child already knew the correct label. 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, Fusaro, & Harris (2009) 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?

Our 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 1c (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 model have a slight preference for the label provided by the informant from the majority (Figure 1c, test).

Deceptive informants

Although all of the informants until now have been helpful, some informants may intend to mislead. How do children reason if they know an informant is deceptive? Mascaro & Sperber (2009) explored this by presenting children with a knowledgeable informant who they were told was deceptive. A piece of candy was secretly placed under one of two cups. The informant (a puppet) looked under both cups 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 which cup the candy

[3] This corresponds to setting β_h and β_k to 0.95, and the strength of these beliefs γ_h and β_h to 5.

was hidden under and the child was asked to guess which cup had the candy. Note that, like previous studies, this is essentially a labeling problem—the puppet is labeling a cup as the cup with the candy in it. The critical question was whether children knew to choose the cup that the puppet did not point to. Indeed, four-year-olds chose the opposite cup approximately 77% of the time, in contrast with their responses at baseline, after watching the puppet look under a different set of cups, 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 using the same helpfulness parameters as the other experiments, but with the knowledgeability parameter set to be high[4] to capture the fact that the puppet, having looking 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 had looked under the cups and always lies by retaining the high prior probability of knowledgeability but changing the prior probability to favor deceptiveness.[5] Figure 1d shows that the model captures the reversal of choice after learning about the informant's deceptive ways. Overall, across the four sets of results, the model provides a close fit to children's behavior ($r = 0.82$).

Developmental Changes

These results suggest that the model captures the behaviour of four-year-olds in epistemic trust problems. However, there are developmental changes from age three to four. In some instances, three-year-olds show the same qualitative behavior as four-year-olds, but less decisively (see Figures 3a, b, c). In others, three-year-olds show sharp qualitative differences (Figure 3d). This raises a question: what causes the changes in behavior between the ages of three and four?

The experiment in which three-year-olds are most different from four-year-olds (Mascaro & Sperber, 2009) relies crucially on using information about an informant's intent. Three-year-olds choose to trust an informant that they are told is a liar, while four-year-olds do not. What explains this difference? One possibility is that younger children have strong default expectations that informants will be helpful (Csibra & Gergely, 2009; Csibra, 2007; Tomasello, Carpenter, Call, Behne, & Moll, 2005). Because strong expectations require more evidence to overcome, three-year-olds may continue to believe the informant is helpful in spite of the experimenter's testimony that the informant "always tells lies."

We test this prediction in two ways. First, we identify a different set of best-fit model parameters for the three-year-olds' data (again, there is one set of parameters for all four studies). If our hypothesis is correct, then the best fitting parameters for intent should be more biased toward helpfulness than those found for the four-year-olds. Consistent with this prediction, the best fitting prior over helpfulness was 0.9 rather than 0.6, and the strength of this bias was 3 rather than 1; this corresponds to a strong bias to believe that the informant will be helpful. As shown in Figure 3, the model qualitatively captures all

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

[5] This corresponds to $\beta_k = 0.9$ and $\beta_h = 0.1$. The strength 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.

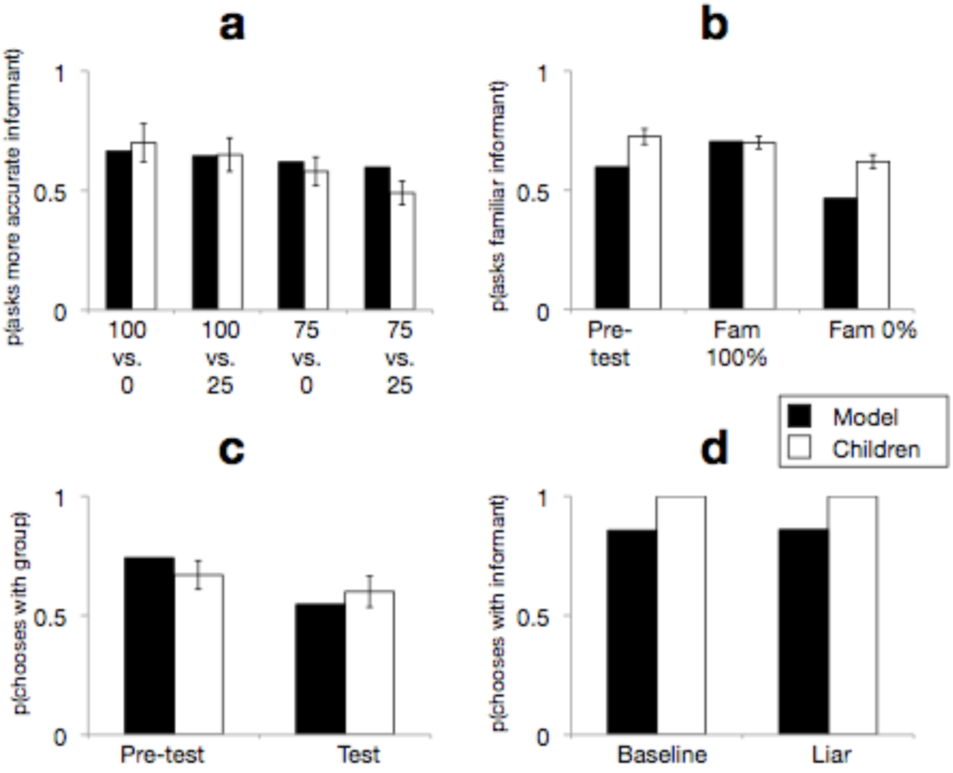

Figure 3. Model predictions and observed results for the performance of three-year-olds in the same epistemic trust tasks as in Figure 1. Across most tasks, three-year-olds' performance is similar to that of four-year-olds, with the exception of inferences about deception (d). In that experiment, three-year-olds 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.

of the experimental findings. Second, we consider the Mascaro & Sperber experiment in particular, shown in Figure 3d. The model makes distinctly different predictions for the three-year-olds than it did for the four-year-olds, suggesting (correctly) that three-year-olds should not behave differently whether told that the informant is a liar or not.

General Discussion

We have presented a model of epistemic trust as inference about the knowledge and intent of informants. Where previous research focused on the importance of knowledge in epistemic trust, we have shown that neither inferences about knowledge nor inferences about helpful intent, alone, can account for the empirical results. We demonstrate that a model that simultaneously makes inferences about knowledge, intent, and the state of the world can predict children's performance in a variety of tasks. Moreover, the model suggests that developmental changes from three to four years of age may stem from differing default assumptions about the helpfulness of an informant.

Although our model suggests that developmental differences in behavior are consistent with different expectations about helpfulness, this is not the only explanation for these differences. For instance, it may be that three-year-olds simply did not understand what "liar" means. Convergent evidence provides some additional support for our interpretation. Corriveau, Meints, & Harris (2009) contrasted children's performance in three conditions: accurate labeler, inaccurate labeler, and neutral (the informant merely drew the child's attention to the object). Children were presented with pairs of conditions, and asked who they would want as an informant in the future. While four-year-old children chose accurate over both neutral and inaccurate, and neutral over inaccurate, three-year-olds did not distinguish between the accurate and the neutral conditions. This result is consistent with our explanation that three-year-olds have a stronger expectation that people will be helpful; if three-year olds have that expectation strongly enough, then evidence of accuracy should not change their beliefs much. Despite this, our results showing that children's behavior is consistent with a developmental change in prior expectations about helpfulness are not definitive, and it is important that future research test this hypothesis more systematically.

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 stimultaneously learn the true state of the world and what 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 three and four years of age may result from changes in children's prior expectations about informants.

Our account of deception, based on the experiments by Mascaro & 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.

In sum, this work provides a novel formal account of epistemic trust as well as an exploration of the implications of epistemic trust for learning. 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 a critical step toward understanding the power of human learning.

Appendix A: Full 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

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

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 learners bias to believe that people are usually helpful, and γ_h, a parameter that describes the strength of that belief:

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

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

$$\begin{aligned} k &\sim \text{Bernoulli}(\theta_k) && (7) \\ \theta_k &\sim \text{Beta}(\gamma_k \beta_k, \gamma_k(1 - \beta_k)) && (8) \end{aligned}$$

and unless otherwise specified in the text, both of the strength parameters γ_h and γ_k were set to 1, and the bias parameters β_h and β_k were treated as free parameters that we fit to the data (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., $\ell = s$), 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:

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

where $\theta = (\theta_h, \theta_k)$ refers to both helpfulness and knowledgeability of the informant, $\beta = (\beta_h, \beta_k)$ refers to the learners biases about helpfulness and knowledgeability, and $\gamma = (\gamma_h, \gamma_k)$ refers to the strengths 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 calculation 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 $P(\ell = s \,|\, E, \beta, \gamma)$ 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 four-year-olds data, we performed a grid search over the values of β_h and β_k. These parameters may take on values between 1 and 0, with 1 on β_h indicating the expectation that people are always helpful and 0 indicating the expectation that people are always deceptive, and the grid search was performed in increments of 0.1. At each pair of values, the sum of squared distance between the observed and predicted value was computed for the Pasquini et al. (2007) and Corriveau, Fusaro, & Harris (2009) data. The results based on the parameter pair with with the smallest mean squared error, $\beta_h = 0.6$ and $\beta_k = 0.6$ (MSE = 0.02, $r = 0.82$), are shown in Figure 1.

To test that the model is not overly complex, and the parameters overfitting the data, we performed cross-validation tests between these two data sets. In these tests, we found the parameters that minimized the mean squared error for one experiment, and tested generalization to the other. The best fitting parameters for the Pasquini data were $\beta_h = 0.5$ and $\beta_k = 0.7$, which achieved an overall MSE of 0.03 and overall correlation of 0.72. The best fitting parameters for the Corriveau et al. data were $\beta_h = 0.5$ and $\beta_k = 0.5$, which achieved an overall MSE of 0.04 and overall correlation of 0.63. Notably, the best fitting parameter values are similar when fit to the subsets and the overall data, and the errors are comparable across the three parameter values. The cross-validation results suggest that the complexity of the model is appropriate for the data, and that the parameters generalize well and are not overfitted.

The Corriveau & Harris (2009a) and Mascaro & Sperber (2009) data were fit separately. Specifically, because Corriveau and Harris used familiar informants who were known to be knowledgeable and helpful in the past, we set these parameters to reflect this prior experience: β_h and β_k were set to 0.95 to capture the expectation that this person is helpful, and the strength of these beliefs was increased to 5. 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, we set $\beta_h = .1$, $\beta_k = .9$, and the strength to be high for both, 5. In both cases, the results were robust across a range of values for the strength parameters.

The procedure for fitting three-year-olds data was similar with minor exceptions. Specifically, to capture three-year-olds' stronger assumptions, the strength parameters were increased from 1 to 3. The knowledgeability parameter was fixed at the same value as for the four-year-olds. We then fit the intent parameter using the grid search method described above to test the hypothesis that three-year-olds inferences are guided by a stronger assumption of helpfulness (the best fitting parameter was 0.9).

References

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

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

Csibra, G. (2007). Teachers in the wild. *Trends in Cognitive Sciences*, *11*, 95–96.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *14*, 148–153.

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.

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

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

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., & Griffiths, T. L. (under review). Rational reasoning in pedagogical contexts. *Manuscript under review*.

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.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*, 675-691.