The Development of Joint Belief-Desire Inferences

Hilary L. Richardson[1] (hlrich@mit.edu), Chris L. Baker[1] (clbaker@mit.edu),
Joshua B. Tenenbaum[1] (jbt@mit.edu), and Rebecca R. Saxe[1,2] (saxe@mit.edu)
Department of Brain and Cognitive Sciences[1] and McGovern Institute for Brain Research[2]
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Human beings infer complex mental states given very little information—a facial expression, a sarcastic tone, or even a simple behavior. Previous work suggests that adults make joint belief and desire inferences based on an actor's path, and that these inferences are well-explained by a Bayesian framework (Baker, Saxe, & Tenenbaum, 2011). We investigate the development of this ability by assessing mental state inferences made by children ages 3-6 after watching a short movie. Our results suggest that young children spontaneously make inductive inferences about desires or preferences, and that the ability to infer belief from behavior develops between ages 3-6, and possibly throughout later childhood. We formulate three computational models that capture the developmental shift between non-representational and representational theory of mind, and show that these models capture qualitative patterns in the children's data.

Keywords: theory of mind, false-belief task, Bayesian inference, cognitive development

Introduction

As we move about the world, our actions are the observable manifestation of unobservable intentions: we act to fulfill our hopes and desires in accordance with our beliefs. Adults understand this intuitively. When a girl exclaims "I'm starving—I'm craving a piece of fruit!" and begins to search extensively for an apple in the kitchen even though a pear is in plain sight, adults can infer that the girl wants to eat a fruit, that she has a preference for apples over pears, and that she has a reasonable degree of belief that there is an apple in the kitchen. Our explanation of the girl's action in terms of inferred beliefs and desires relies on a Theory of Mind (ToM): we understand that agents have a working representation of the world that may or may not reflect reality, that this representation is influenced by perceptual access and priors, and that this representation is the basis for subsequent behavior (Gopnik & Wellman, 1992).

If the girl believes that apples are in the fruit basket, we confidently predict she will look for one there, even if we know that the apples are actually in the cupboard. This ability is assessed by the famous "False-Belief task[1]" (Wimmer & Perner, 1983), on which children typically transition from failure to success between the ages of three and five (Wellman, Cross, & Watson, 2001). Most prior work studying the development of representational ToM has used versions of the False-Belief task to ask children to predict an agent's behavior, given the agent's previously established beliefs and desires. In contrast, there has been less work studying how children solve the *inverse problem*: inferring beliefs and preferences from an observed behavior. Given the girl's extensive search for a fruit in the kitchen, how do we simultaneously infer her beliefs and preferences? Both kinds of judgments—predicting actions given beliefs and desires, and inferring beliefs and desires given actions—tap similar ToM reasoning abilities. This paper aims to test whether the development of the ability to make ToM inferences parallels the transition to understanding false beliefs, and to provide evidence for a formal account of the knowledge supporting both ToM abilities.

The ability to solve this inverse problem is analogous to solving one equation with two unknowns; our natural ability to consider context, weigh in with priors, and make rational inferences enables us to come up with a good guess on questions we would otherwise not be able to answer. Studying this ability in the social domain illustrates the power of ToM to go "beyond the data" and infer multiple implicit mental states from just one observed action. Prior work by Baker et al. (2011) presented adult participants with an inverse mental-state-inference task and showed that adult mental state inferences are well-explained within a rational probabilistic inference framework. Here, we use an analogous paradigm to measure spontaneous mental state inferences made by children 3-6 years of age and assess which observed behaviors prompt mental state inference. By doing so, we are measuring children's expectation that all parts of an action should have a sufficient explanation in terms of mental states.

If this inferential ability develops in parallel with the ability to predict behavior given a mental state, we would expect a similar shift in performance between the ages of three to five on inverse problems that require mental state inference. On the other hand, it is possible that the ability to infer mental states from sparse information develops later in life. This process not only requires the ability to take the perspective of another and maintain multiple representations of the world, but it also requires that the viewer spontaneously seeks to understand observed actions in terms of underlying beliefs and desires.

In our experiment, children watched a short 3D animation of a hungry bunny navigating a world to find and eat one of three different fruits. The bunny can take one of three paths: (1) pass the nearest, visible fruit to check around a wall to choose the fruit there (2) take a direct path to the nearest,

[1] "Explicit" vs. "implicit": Surprisingly, infants succeed in looking-time paradigms tapping analogous notions of perceptual access and false belief representation (Onishi & Baillargeon, 2005).

visible fruit (3) take the longest path, passing the nearest fruit to check around the wall (in sight of the farther fruit), and then turning back to choose the near fruit. These three paths suggest different orders and degrees of preference for the three fruits, according to both our intuition and three computational models. In particular, the third path can only be fully interpreted as rational if children infer that the bunny was "looking for" the missing fruit. We analyze how many children inferred the bunny's fruit preference, as a function of age and condition (i.e., the bunny's path). In doing so, we believe we are assessing a sophisticated aspect of ToM reasoning: the ability to make rich inverse inferences from limited data.

Methods

Participants

143 children were recruited from a local children's museum to participate in the study. Out of these, 103 were included in the final sample (70 females, 3-4yo group: M=4.02, SD=0.62; 5-6yo group: M=5.79, SD=0.56). 40 participants were excluded; 28 for answering at least one of four memory and control questions incorrectly, 5 for parent or sibling interference, 4 for not answering all of the questions, and 3 for experimenter error.

55 adults were recruited from an MIT human subjects listserv. Out of these, 54 were included in the final sample (30 females, M=25.95, SD=6.14). 1 participant was excluded for not following task instructions.

Stimuli and Design

3D animated movies were created using Alice 2.2 programming software (http://www.alice.org). Stimuli are available at: http://saxelab.mit.edu/bunny/. Each participant watched one movie two times followed by a short ending.

Movie Introduction Each movie begins with a bunny standing on a green platform. A brown wall divides the platform—this wall reaches above the bunny's eye level, obstructing his view of the other side. There is a tree with three different fruit on the bunny's side of the wall. The fruit varied in shape and color (yellow, red, and orange), and position of the fruit was counterbalanced across children. At the beginning of the movie, the bunny waves and says "Hello!" He then turns to the tree, points at the three fruits (ambiguously), and says, "I'm hungry, I want that one!" He attempts to reach the fruit, and eventually succeeds in knocking all three fruits off the tree simultaneously. While the bunny is still facing the tree, the three fruit fall down and roll away—one lands in plain sight of the bunny (Fruit 3), but the other two roll to the other side of the wall. One of these fruits stays on the other side of the wall (Fruit 2), and the other (Fruit 1) falls off of the platform. In the movies viewed by children, Fruit 1 rolled out of sight. The bunny turns to get his fruit, sees Fruit 3, and takes one of three paths: Check Stay (CS), No Check (NC), or Check Turn (CT) (see below for detailed description).

This introduction was created to allow viewers to understand the bunny's initial belief state about the world. Before he embarks on one of three paths to find a fruit, participants are provided with evidence that the bunny knows that the three fruit exist, and knows that they rolled away. They are also prompted to understand that the bunny is hungry and has a preference for one fruit over the others. Finally, participants know that the bunny has initial perceptual access to Fruit 3, and some degree of belief that the other fruits might be on the other side of the wall. These priors were similarly built into our computational models.

Between the two movie viewings, participants saw a black screen with the words "Let's watch one more time!" for 3 seconds. The experimenter read this text out loud to child participants at this time.

Paths The bunny's paths were designed to evoke inferences about the bunny's preferences and beliefs. Below we describe each path, and note in italics the inferences that each path elicits. While we believe these inferences are intuitive, we also characterize them through three formal computational models (see Computational Models section).

Check Stay (CS) The bunny passes Fruit 3, walks to the other side of the wall, and chooses Fruit 2 (see Fig. 1). *Predicted Inferences: all age groups will infer that Fruit 2 is the bunny's favorite fruit. Passing Fruit 3 is the strongest evidence from which to infer that Fruit 3 is the least favorite fruit. Fruit 1 is underspecified.*

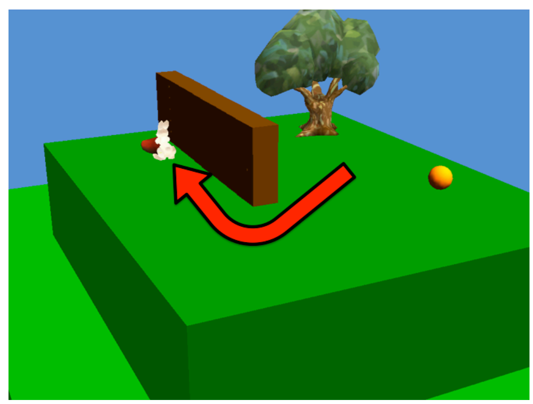

Figure 1: Final frame of Check Stay Condition (red arrow shows path).

No Check (NC) The bunny walks directly to Fruit 3 (see Fig. 2). *Predicted Inferences: all age groups will infer that Fruit 3 is the favorite fruit. There is no distinguishing evidence for preference order between Fruit 1 and Fruit 2, so the bunny's least favorite fruit remains ambiguous.*

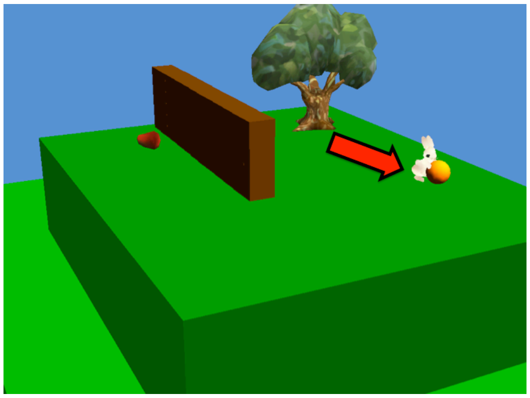

Figure 2: Final frame of No Check Condition (red arrow shows path).

Check Turn (CT) The bunny passes Fruit 3, walks to the other side of the wall, sees Fruit 2, and goes back to Fruit 3 (see Fig. 3). *Predicted Inferences: Participants who attribute initial uncertainty about the locations of the occluded fruits to the bunny will infer the correct preference order: the bunny's favorite fruit is Fruit 1, and his least favorite fruit is Fruit 2. Participants who do not consider the bunny's beliefs will infer that the bunny's favorite fruit is Fruit 3 and his least favorite fruit is Fruit 2.*

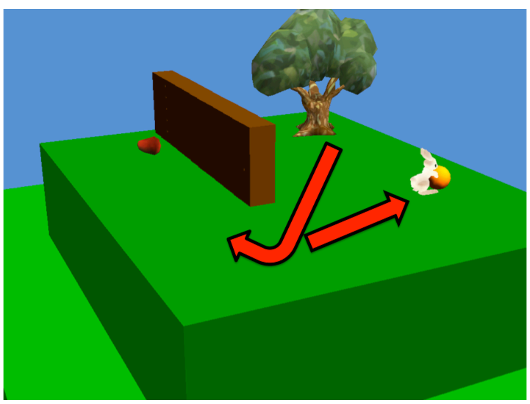

Figure 3: Final frame of Check Turn Condition (red arrow shows path).

Movie Ending After children viewed both the introduction and path twice, they viewed a short ending. This ending shows all three fruit roll within sight and reach of the bunny, making it clear that this time, the bunny can choose any of the fruits. All participants were shown a laminated picture of the final frame of this ending, which shows the bunny equidistant from all three fruit. This picture enabled children to respond to questions by pointing, and was placed in front of adults as they filled out the short questionnaire.

All participants sat approximately 12-18" from the movie, which was played on a 15" MacBook Pro. Each participant watched two trials of one condition for a total watching time of 3:11 minutes (CT condition), 2:48 minutes (CS condition), or 2:25 minutes (NC condition).

Adult participants were asked to rank how much the bunny liked each of the fruits on a Likert scale from 1-7 (Not Much-Very Much). They were also asked to explain their rankings ("How do you know?"). All responses were recorded by the participant using pencil and paper.

Child participants were asked two control questions, four test questions, and two memory check questions after watching the movie. The first control question asked the child to identify each fruit by pointing to it on the laminated picture; the order of fruits was counterbalanced across children. The second control question asked, "At the beginning of the movie, how many fruit did the bunny see in the tree?" Children were then asked "Which fruit do you think is the bunny's favorite?" and "How do you know?" Next, children were asked "Which fruit do you think is the bunny's least favorite—which one does he not like?" and "How do you know?" Finally, children were asked "During the movie, which fruit fell off and rolled away?" and "During the movie, which fruit did the bunny pick?" The control questions provided reasonable exclusion criteria and enabled us to confidently report that children were engaged by the movie and remembered its content, while the test questions provided us with sufficient information regarding preference understanding (which fruit was favorite, neutral, and least favorite). Most children responded to the favorite and least favorite questions by pointing to the laminated picture (rather than responding verbally). Child testing sessions were recorded using a video camera. Parents were usually in the testing room with the child participants, but were asked to remain quiet during the experiment.

Computational Models

We model children and adults' mental state inferences as a kind of probabilistic causal reasoning. One way to approach inverse problems is through Bayesian inference, which describes the process by which people generate and test hypotheses based on their expectations and evidence. The probability of a hypothesized mental state H, given observed evidence E, denoted P(H|E), depends on P(E|H), the likelihood of the evidence, given the hypothesis, and P(H), the prior probability that the hypothesis is true, according to Bayes' rule:

$$P(H \mid E) \propto P(E \mid H)P(H).$$

Posed in the context of this study: given the path that the bunny takes and the fruit that he chooses (E), what were his initial beliefs and desires (H)?

To capture the development of the ability to represent and infer mental states, we formulated three models: Outcome-Based (OB), Desire Theorist (DT), and Bayesian ToM (BToM). The Outcome-Based model assigns full preference to the fruit that the bunny chooses; it is not influenced by the bunny's beliefs or the path that he takes. The DT model is based on the Copy Theorist model of Goodman, et al. (2006). DT is a non-mentalistic model that represents how an agent's desires and the world state—but not the agent's beliefs—cause its actions, via the principle of rational action (Gergely et al., 1995). The DT model can infer straightforward desires that do not depend on the actor's belief state. We expect this model to explain mental state inferences that derive mainly from which fruit the bunny chose. The third model is based on the BToM framework proposed by Baker et al. (2011), which models a mentalistic theory. BToM incorporates a principle of rational belief: the assumption that agents maintain beliefs about the world that are consistent with physical laws and depend on their perceptual access to the world (i.e., visual access to different fruits). This enables simultaneous inference of agents' beliefs and desires, given observations of their behavior.

The DT and BToM models assume that the bunny's actions are guided by his degree of desire (subjective value) for each fruit, which trades off against his expected costs to reach them. The bunny incurs costs for effort (quantified by the number of steps taken per path: 1.5 steps for NC, 2.75 for CT, and 3 for CS) and time (an additional cost of .25 for checking around the wall). The bunny's favorite fruit is assigned a value of 15, the second-favorite fruit a value of 5, and the least favorite fruit a value of 1. This desire scale reflects a strong preference for one fruit and is calibrated to the spatial scale of the environment; changing these approximate values does not alter the trends observed. The OB model also assumes that the bunny's actions are guided by desire, but does not assign incorporate costs or rewards; the chosen fruit is considered the bunny's favorite, and the two unchosen fruits are equally least favorite.

The BToM model attributes initial uncertainty to the bunny about the locations of the two non-visible fruit: both, one, or neither of the two fruits may be available behind the wall. In the BToM model, the bunny's beliefs are updated when he moves to the other side of the wall. The Desire Theorist model, on the other hand, assumes the bunny's actions depend only on the true world state. For BToM, costs incurred to check around the wall are rational if the bunny desires one of the two non-visible fruits and believes it could be there; in the DT model, there is no explanation of why the bunny incurred these costs.

This key difference is most evident in the Check Turn condition, which allows us to assess whether three/four-year olds, five/six-year-olds, and adults incorporate information about uncertainty, planning, and belief updating into their desire inferences. If so, their responses should be better predicted by the BToM model than the DT model. This difference is comparable to the performance shift on the False-Belief task between children who are three and five years old; three-year-old children refer to salient outcomes, actions, and desires, while five-year-old children take beliefs into consideration.

Data Analysis

Child responses to the four control and memory questions, two preference questions, and two explanation questions were transcribed from the recording of the testing session. From these data we recorded the proportion of participants in each age group (3-4yo, 5-6yo, and adults) who reported each fruit (1,2, and 3) as "favorite" and "least favorite" in each condition (CS, NC, CT). We binarized adult participant data; the fruit that was assigned the highest number on the Likert scale was coded as the favorite, and the fruit assigned the lowest number was coded as the least favorite. Whereas all participants picked one fruit as the favorite for each condition, participants often reported two fruits as least favorite. In this case, each fruit received half the weight assigned to a single favorite or least favorite fruit (.5).

We used logistic regression to test for main effects of age (a continuous variable), condition (CS, NC, CT), and age by condition interactions on favorite and least favorite fruit choice. We used a Bonferroni correction (n=3) for multiple comparisons.

To compare behavioral judgments to the three models, we separated participants into three groups: younger children (age 3-4 years, n=56), older children (age 5-6 years, n=46), and adults (n=54). We calculated the probability of choosing each fruit (3) as favorite or least favorite (2) in each condition (3). We compared the resulting 18 values for each group to the corresponding predictions from each model, using a Pearson's correlation.

Results

Check Stay (CS) Condition

This condition is a good measure of spontaneous understanding of preference; the bunny picks Fruit 2 after explicitly passing Fruit 3. We expected that children of all ages as well as adults would correctly identify Fruit 2 as the favorite, as predicted by all three models. We found a positive main effect of the CS condition on choosing Fruit 2 as the favorite fruit ($p<0.001$), and a negative effect of the CS condition on choosing Fruit 2 as the least favorite fruit ($p=0.043$). There were no significant effects of age on fruit choice in this condition. Fig. 4 shows that participant judgments were well predicted by the DT and BToM models across age.

No Check (NC) Condition

This condition is an interesting measure of how participants understand preference when there is less evidence available. The bunny approaches the only visible (and closest) fruit, Fruit 3, providing weaker evidence for his preference; his choice may reflect efficiency or lack of options, rather than a strong preference. Again, all groups successfully picked Fruit 3 as the most likely favorite. We found a positive main effect of the NC behavior on choosing Fruit 3 as the favorite ($p<0.001$), and a negative main effect of the NC behavior on choosing Fruit 3 as the least favorite ($p=0.020$). There was no significant difference between age groups.

We were specifically interested in whether observers were sensitive to the weaker evidence in NC compared to CS paths; if so, Fruit 1 should be more likely to be chosen as least favorite in the NC condition than in the CS condition (because in the CS condition, the bunny explicitly avoided Fruit 3). Only the BToM model showed this qualitative pattern. Although this difference was not significant in any age group individually, combining across age groups did

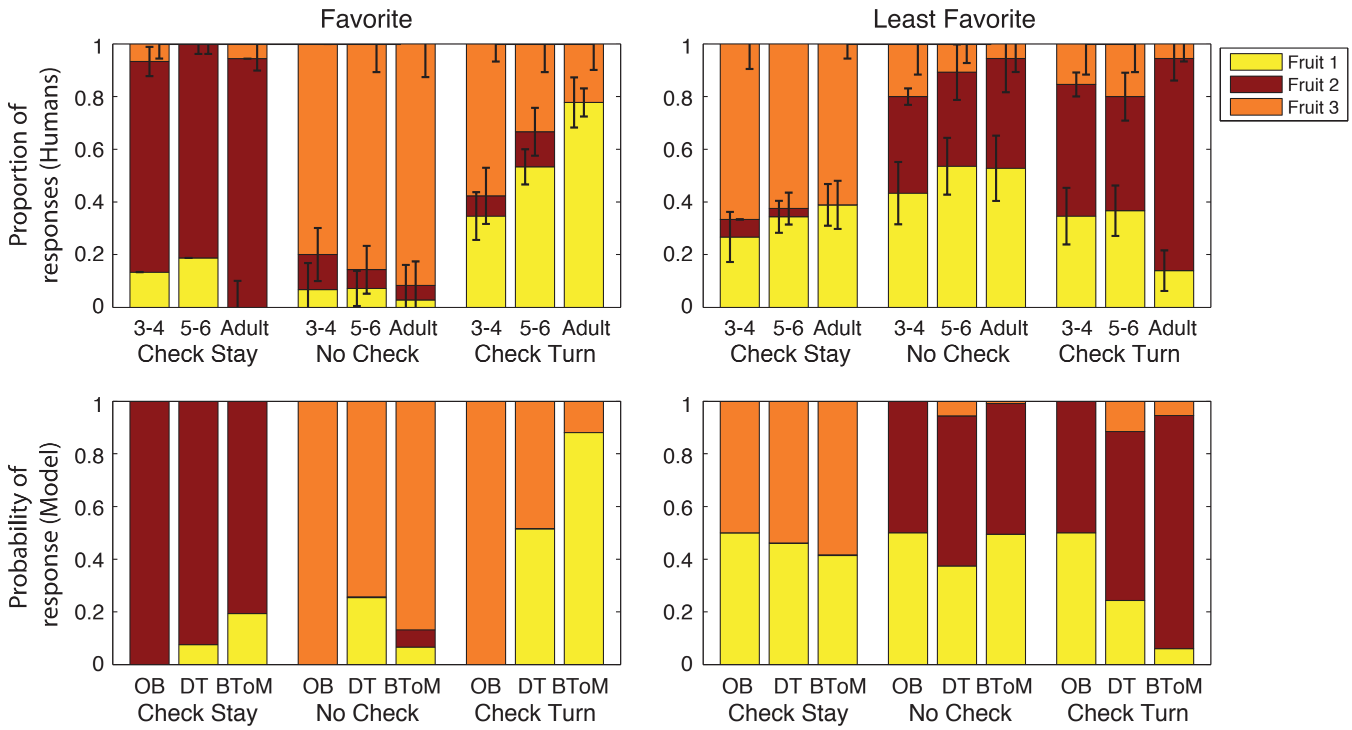

Figure 4: Results of "Favorite" and "Least Favorite" judgments across conditions and age groups. The bottom row compares the results of the OB, DT, and BToM models to human judgments, matching the qualitative developmental shift between younger and older participants.

reveal the predicted pattern (25/47 chose Fruit 1 as least favorite in NC, compared to 16/49 in CS, $p<0.05$, Fisher's exact test).

Check Turn (CT) Condition

This condition tests spontaneous inference of both the beliefs and preferences of the agent. It requires participants to consider the agent's actions in terms of preferences (which fruit did he really want?) and counterfactuals (what was he looking for behind the wall, that he could have seen but did not?) rather than outcome (which fruit did he pick?). Given the well-documented shift from failure to success on False-Belief tasks, we expected 3-4 year olds to perform significantly worse than 5-6 year olds and adults on the "favorite fruit" question.

We found a positive age by condition interaction on choosing Fruit 1 as the favorite fruit ($p=0.029$) and a significant negative age by condition effect on choosing Fruit 1 as least favorite ($p=0.021$). These results show that as participants got older, their judgments slowly came to better resemble the predictions of the BToM model. The salient fact that the bunny chose Fruit 3 was difficult for younger participants to ignore, as suggested by the OB and DT model predictions. As a result, there was also a positive main effect of the CT condition on choosing Fruit 3 as the favorite ($p=0.001$) and a negative main effect of condition on choosing Fruit 3 as least favorite ($p=0.020$).

Model Comparison

Combining across the three conditions, we compared the responses of each age group to the three models. Judgments made by 3-4 year old children were most strongly correlated with the predictions of the OB model (r=.902), followed by the DT model (r=.898), and least with the BToM model (r=.719). In contrast, judgments made by 5-6 year old children correlated most strongly with predictions made by the DT model (r=.898) followed by the BToM model (r=.812), and correlated least with the OB model (r=.786). Adult preference predictions were most strongly correlated with the BToM model (r=.943), followed by the DT model (r=.920), and least correlated with the OB model (r=.678). These data support the idea that across development, people increasingly incorporate spontaneously attributed beliefs and desires into their understanding of agents' behavior.

Discussion

In this study we used a novel ToM task to examine the development of the ability to spontaneously attribute mental states in order to understand an agent's observed actions, i.e., to solve inverse social inference problems. Our results suggest that this ability emerges gradually, with performance continuing to improve after age 5-6 years.

In particular, on the critical Check Turn condition a sufficient explanation of the bunny's behavior requires children to infer that the bunny checked behind the wall because he was looking for (and preferred) the missing Fruit 1. This inference depends on recognizing that the bunny initially didn't know which fruit was behind the wall, and that he couldn't have been looking for Fruit 2, because he subsequently turned back to Fruit 3 (which he initially bypassed). Observers would only make this complex inference if they sought a sufficient rational explanation of the bunny's whole path. Most 3-4 year old children effectively ignored the "checking" path, inferring that the bunny prefers whichever fruit he chooses in the end. This behavior mimicked the prediction of our Outcome-Based and Desire Theorist models. Five and six year olds were somewhat more likely to take the checking path into consideration, recognizing Fruit 1 as the favorite, and adults were even more likely to do so. This created a gradual increase with age in the match between participants' choices and the predictions of the Bayesian Theory of Mind model.

One limitation of the current experiment is that we did not test children older than 5-6 years, so we cannot say whether children's performance on these tasks reaches adult levels by age 7-9 years, or whether there is extended development through adolescence. Another limitation is that children's choices for the "least favorite" fruit were frequently ambiguous (both fruits not identified as favorite were often considered least favorite). We are currently working on an extension in which children provide a full ordering of the bunny's preference for all three fruits.

The current results are amenable to multiple interpretations. First, children may be becoming more sophisticated and adept at thinking about other minds. Simultaneously inferring a belief and a desire may require a more robust ToM capacity than using beliefs and desires to predict actions. If so, these results may be related to the observation that children's ToM brain regions also become increasingly selective after age 5-6 years (Gweon, Dodell-Feder, Bedny, & Saxe, in press). Another possibility is that over development, children become more committed to the idea that agent's actions are rational, and require sufficient rational explanations. Thus, younger children may be less likely to view others' actions as efficient paths toward their goals, and so the bunny's deviant path may not seem to require any explanation. As children expect more efficiency from others, the deviant path may become more salient, and demand an explanation. Finally, a third possibility is that children's ability to focus on and interpret the bunny's path, in addition to his (highly salient) final position, depends not on ToM development per se, but on unmasking of prior competence by the development of executive function (Carlson, Moses, & Breton, 2002; Baillargeon, Scott, & He, 2010).

Another key question is how infants will perform on this task. Recent evidence from looking time suggests that infants have a nascent ToM, and can update representations of agents' beliefs and desires given their perceptual access, in change-in-location and appearance-reality tasks (for a review, see Baillargeon et al., 2010). We believe that if infants make simultaneous inferences about beliefs and desires in the current paradigm, it would be particularly strong evidence for a mentalistic account of infant ToM abilities.

In sum, the current study is an initial step toward clarifying how we develop the ability to make joint belief-desire inferences in order to understand other minds. It contributes to the current literature on ToM development, suggesting that children develop this ability in parallel with other mentalistic reasoning abilities between the ages of three and five, and could serve as a launching point for future work studying rich social inferences made by infants.

Acknowledgements The authors were supported by the Packard Foundation. We thank members of the Saxe and Schulz labs for thoughtful discussion and feedback and Mika Asaba for help with data collection. Thanks also to the Children's Museum of Boston and participating families.

References

Baillargeon, R., Scott, R.M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14,* 110-118.

Baker, C.L., Saxe, R.R., & Tenenbaum, J.B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469-2474).

Carlson, S.M., Moses, L.J., & Breton, C. (2002). How specific is the relation between executive functioning and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*, 73-92.

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193.

Goodman, N.D., Baker, C.L., Bonawitz, E.B., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J.B. (2006). Intuitive Theories of Mind: A Rational Approach to False Belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1382-1387).

Gopnik, A. and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7,* 145-171.

Gweon, H., Dodell-Feder, D., Bedny, M., and Saxe, R. (in press). Theory of Mind performance in children correlates with functional specialization of brain regions recruited for thinking about thoughts. *Child Development*.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development* 72, 655-684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.