Beliefs about desires: Children's understanding of how knowledge and preference influence choice.

Julian Jara-Ettinger, Emily Lydic, Joshua B. Tenenbaum, & Laura E. Schulz
(jjara, emilydic, jbt, lschulz @ mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Abstract

Knowledgeable agents always choose what they like best, thus revealing their preferences. But naïve agents only choose what they believe they like best, and may end up disliking their choice. As such, sensitivity to an agent's prior experience is critical for interpreting their behavior. Here we show that four- and five-year-olds expect knowledgeable agents, as compared to naïve agents, to have stable choices that lead to higher rewards (Experiments 1 and 2). Additionally, we show that four- and five-year-olds can infer which of two agents is naïve given information about the rewards they obtained and the stability of their choices (Experiments 3 and 4). These results show that young children understand that beliefs and desires are interconnected and that, in addition to having uncertainty about the world, agents can also be uncertain about their own desires.

Keywords: Social Cognition; Theory of Mind.

Introduction

Humans have strikingly sophisticated social skills (Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007). We understand that behind other people's actions lies a rich mental life. Although we cannot directly observe others' mental states, we have an intuitive theory that enables us to infer them. This understanding emerges early in life (Luo, 2011; Onishi & Baillargeon, 2005; Woodward, Sommerville, Guajardo, 2001) and enables us to make rich and powerful inferences in adulthood (Baker, Saxe, & Tenenbaum, 2009, 2011; Jara-Ettinger, Baker, & Tenenbaum, 2012).

This intuitive theory, called a Theory of Mind (ToM), connects information about people's desires and beliefs with their actions (e.g., Dennett, 1989; Wellman, 1990; Gopnik & Meltzoff, 1997). If Sally *wants* a cookie and *believes* there are cookies inside the cookie jar, we can predict that Sally will walk toward the cookie jar and take a cookie out. Having this causal theory also enables us to infer Sally's beliefs and desires from information about her actions. If Sally takes a cookie from the cookie jar and eats it, we can infer that she wanted a cookie and that she (correctly) believed that she would find one in the cookie jar. If instead, Sally walks away empty handed after peeking into the cookie jar, we may infer she had a false belief about the cookie jar's contents, and thus failed to fulfill her desire. Recent computational work formalizing the ways that beliefs and desires jointly determine an agent's actions can predict human judgments with quantitative precision (Baker et al., 2009, 2011; Jara-Ettinger, Baker, & Tenenbaum, 2012).

All of these accounts treat beliefs and desires as independent variables, and characterize beliefs as representing content about the world. However, in addition to having beliefs about the world, agents have beliefs about what will fulfill their desires. Consider again the simple case of watching Sally get a cookie from the jar. We might infer that Sally likes cookies and that she would get a cookie again if she found herself in the same situation. However, this inference assumes that Sally not only knew there were cookies in the jar, but also knew that she liked cookies. If we knew that Sally had never tried a cookie before, we might not be so fast to assume that Sally will eat cookies in the future.

Cookies, of course, are almost universally familiar and universally liked, but in novel contexts, inferences that depend on an agent's beliefs about her desires are both commonplace and critical for social cognition. Imagine for instance, that you are watching your friend buy food at a market. Usually, her choices reveal what she likes. However, if she's in a foreign country and has never tasted the food before, her choices only reveal her best guess; they may not tell us anything about her stable, long-term preferences. For us to know what someone likes, she has to know it herself first. We can draw comparable inferences in reverse. If you see your friend try a chocolate from her box and then change her mind and choose a different one, you might infer that she was initially naïve, unsure, or wrong, about what was inside the chocolate. In these cases, the instability of the agent's preferences is indicative of the initial uncertainty of her beliefs about her utilities.

Despite extensive past work on the development of theory of mind, to our knowledge no work has examined children's understanding of how uncertainty about one's own desires influences behavior or at their ability to infer knowledge or ignorance about utilities using information about agents' actions. To the degree that researchers have looked at children's inferences regarding how agents change their mind with evidence, they have focused primarily on issues related to epistemic access: canonically, whether the agent does or does see where a desired object has been placed (e.g., Wimmer & Perner, 1983). However, an agent may also be aware or unaware of the value of a putatively desired object. Thus, the degree to which the agent's initial estimates are stable depends on how much the agent knows initially.

In these studies we investigate children's understanding of how agents' uncertainty about their desires relates to the expected outcome of their goals, and to the stability of their behavior. Figure 1 shows a graphical display of the experiments. In Experiment 1 we ask whether children understand that knowledgeable agents are more likely than naïve agents to take actions that lead to a high reward. In Experiment 2, we ask whether children understand that knowledgeable agents are more likely than naïve agents to make decisions that are stable over time. Experiments 3 and 4 examine the inverse questions: In Experiment 3, we ask if children believe that agents who obtain a high reward are more likely to have been knowledgeable, and in Experiment 4, we ask if children believe that agents who make more stable choices are more likely to have been more knowledgeable. The current study goes beyond merely representing agents' beliefs; instead children must understand that different agents might perform the same actions with the same beliefs about the world and yet interpret the experience differently. Because children's ability to reason explicitly about agents' mistaken beliefs emerges between ages four and five (Wellman, Cross, & Watson, 2001) here, we focus on four- and five-year-olds.

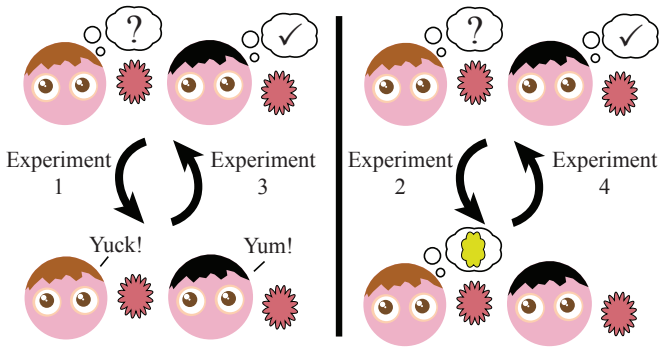

Figure 1: In Experiments 1 and 2 children saw a knowledgeable and a naïve agent make the same choice. Children were asked to infer which puppet said "yum" and which puppet said "yuck" (Experiment 1), or which puppet changed their mind (Experiment 2). Experiments 3 and 4 tested inferences in the reverse direction.

Experiment 1

In Experiment 1 we test if children understand that agents' choices are affected by their estimate of the expected rewards of their actions, and thus that knowledgeable agents are more likely than naïve agents to accrue high actual rewards. Children were introduced to two puppets who had been given a choice between two types of fruits. One puppet was knowledgeable and had tasted both fruits before; one puppet was naïve and had not. Both puppets chose the same fruit. One puppet tasted it and said "Yum!" and one puppet tasted it and said "Yuck!" We looked at whether children inferred that the knowledgeable puppet was more likely to say "Yum!"

Methods

Participants 16 participants (mean age (SD): 5.09 years (195 days), range 4.13-5.89 years) were recruited at an urban children's museum. One additional participant was recruited but not included in the study because he failed to respond the inclusion question correctly (See Procedure).

Stimuli The stimuli consisted of two pairs of gender-matched puppets, and picture cutouts of two fruits: Rambutans, and African cucumbers.

Procedure Participants were tested individually in a quiet room in a children's museum. The child and the experimenter sat on opposite sides of a small table. The experimenter first introduced the cutout pictures of the rambutans and the African cucumbers and placed four pictures of each fruit on the table with each kind of fruit in its own pile. Next, the experimenter introduced the two puppets by name ("Anne" and "Sally", or "Arnold" and "Bob", depending on the participant's gender. The puppets were matched with the participant's gender to ensure gender biases did not influence our task). The experimenter then explained that "Sally has never seen these fruits before and she doesn't know what they taste like" while "Anne knows all about these fruits. She knows what they taste like." (Knowledgeable puppet and introduction order were counterbalanced). Next, the experimenter told the participant "Earlier today, we told our friends they had to pick one fruit to each, and both of our friends picked a rambutan." (Actual fruit counterbalanced). Next, the experimenter placed a picture of a rambutan in front of each puppet and explained, "Both of our friends took a bite of their fruit and one of them said 'Yum!' and one of them said 'Yuck'!" Participants were then asked an inclusion question to ensure the child remembered the critical information: "Can you tell me, which of our friends has not tasted the fruits before? And which one of our friends has not tasted these fruits before?" Finally, participants were asked which puppet said "Yum!" and which puppet said "Yuck!"

Results and Discussion

Children who failed to respond correctly to the inclusion question were excluded from analysis and replaced (n = 1). Results were coded for adherence to the script by a coder blind to the child's response to the test question (no participants were dropped due to experimenter error). Videotapes were then coded to record the child's response to the test question. Children were coded as answering correctly if they indicated that the knowledgeable puppet had said "Yum." Of the sixteen children who responded to the inclusion question correctly, 100% responded correctly to the test question (95% CI: 82.93%-100%. See Figure 2).[1]

[1] Due to conceptual issues with Null hypothesis significance testing (e.g., Cohen, 1994), and a recent proposal to move the field towards better standards for analyzing data (Cumming, 2013), we present confidence intervals as our main method of analysis.

Note that if children believe that an agent's choices always reflect her preferences then children should have expected both puppets to say "Yum!" That is, if children recovered agents' desires only from information about the agents' actions and beliefs about the state of the world, then children should have responded at chance in this context. Both agents knew they had a choice of the two fruits and both agents made the same choice. Children instead recognized that the knowledgeable agent would be more likely to like the chosen fruit, which suggests that children understand that agents choose the options with the highest expected rewards and that a naive agent's choices may be governed by an inaccurate estimate of the actual reward.

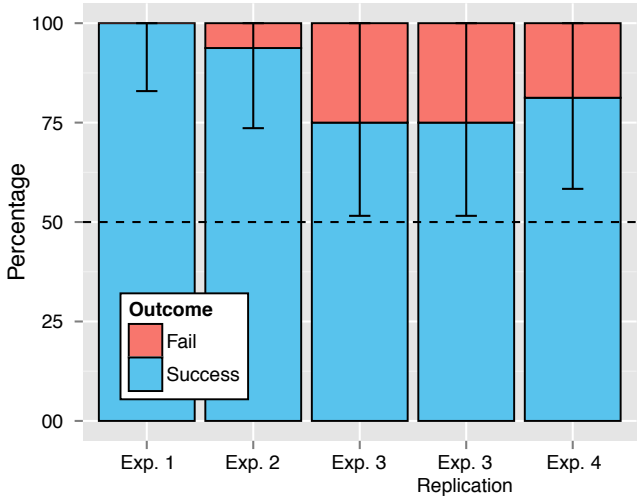

Figure 2: Results from all experiments. The x-axis shows each experiment and the y-axis shows the distribution of children's choices (color coded). The dashed horizontal line represents expected chance performance and the vertical solid lines show 95% confidence intervals.

Experiment 2

The results from Experiment 1 suggest that four and five-year-olds understand that, relative to knowledgeable agents, naïve agents are more likely to make unrewarding choices. When this happens, naïve agents will most likely reconsider their choices. However, naïve agents may reconsider their choice even when they obtain a positive reward, especially if the reward is lower than they expected. Thus, in general, naïve agents are more likely to have unstable choices, when compared to knowledgeable agents. In Experiment 2 we test if, in the absence of knowledge about the rewards, children have expectations about the stability of the choices of knowledgeable and naïve agents.

Methods

Participants 16 participants (mean age (SD): 5.16 years (241 days), range 4.01-5.96 years) were recruited at an urban children's museum. One additional participant was recruited but excluded from the study and replaced because he declined to complete the task.

Stimuli The stimuli were identical to those used in Experiment 1.

Procedure The procedure was identical to the procedure in Experiment 1 except as follows: Neither puppet said "Yum!" or "Yuck!" Instead, after the puppets had tasted the fruits, the experimenter said, "Both of our friends took a bite from their fruit and one of them changed her mind and decided she wanted to eat a different fruit". For the test question, participants were asked, "Which one of our friends changed his/her mind?"

Results and Discussion

As in Experiment 1, children who failed to respond to the inclusion question were excluded from the study. (No participants were dropped on these grounds.) Videos were first coded for adherence to script by a coder blind to the child's test response (no participants were dropped due to experimenter error), and later to record the child's answer to the test question. Children were coded as responding correctly if they indicated that the naïve agent was the one who had changed her mind. Fifteen of the 16 children responded correctly to the test question (93.75%; 95% CI: 73.60%-100%. See Figure 2).

Note that in contrast to Experiment 1, participants in Experiment 2 never obtained any information about the outcome of each puppet's choice. Thus, it was possible that either or both puppets had liked or disliked their chosen fruit. Nevertheless, children were able to infer that naïve agents are more likely to make unstable choices. Furthermore, in this experiment we used a neutral dependent measure, thus ensuring that participants couldn't succeed by grouping together two features with a positive valence (such as knowledge and "yumminess" in Experiment 1). Together with Experiment 1, these results suggest that four and five year-olds understand that relative to knowledgeable agents, naïve agents are more likely to make choices that lead to low rewards, and thus that their choices are less likely to be stable over time.

Experiments 1 and 2 suggest that children have different expectations about the rewards that knowledgeable and naïve agents obtain and about the stability of their actions. In Experiments 3 and 4 we ask if children can reverse these inferences, and infer an agent's prior knowledge based on whether the agent succeeds in obtaining high rewards (Experiment 3), and whether the agent's choices are stable (Experiment 4).

Experiment 3

In Experiment 3 we invert the question asked in Experiment 1. Here we ask if children believe that agents

who make choices that result in low rewards are more likely to have been naïve prior to making their choice. Children watched two puppets pick the same fruit to eat. After learning that one puppet said "Yum!" and the other puppet said "Yuck!" children were asked to decide which puppet had not tasted the fruits before.

Methods

Participants 32 participants (mean age (SD): 5.12 years (194 days), range 4.12-5.98 years) were recruited at an urban children's museum. Sixteen participants were recruited for the original experiment, and sixteen additional participants were recruited to conduct a replication (see Results). Four additional participants were recruited in the original experiment but excluded from analysis and replaced for failing the inclusion question (n = 1), declining to complete the experiment (n = 1), and declining to answer the test question (n =2). One additional participant was recruited in the replication experiment but excluded from analysis because he declined to answer the test question. See Results.

Stimuli The stimuli were identical to those used in Experiment 1.

Procedure Participants were tested individually in a quiet room. As in Experiment 1, the experimenter introduced the two puppets and the fruits and explained that each puppet chose a fruit to eat. The section from Experiment 1 in which where the experimenter explained that one puppet was more knowledgeable than the other was omitted. After both puppets chose a fruit, the experimenter said "Anne and Sally (or Arnold and Bob) both took a bite from their rambutans (or African cucumbers). Anne said 'Yum!' Sally said 'Yuck!'" Next, the experimenter said, "But guess what? One of our friends didn't know what rambutans tasted like until today." Children were then asked to remember which puppet had said "Yum!" and which puppet had said "Yuck!" For the test question, the experimenter asked, "Can you tell me, which one of our friends didn't know what rambutans tasted like until today?" (Actual fruits counterbalanced throughout.) The replication experiment had the same procedure as the original experiment with the exception that the inclusion question was asked immediately after the puppets tasted the fruit, thus making the last part of the experiment more fluent.

Results and Discussion

Results were coded as in Experiments 1 and 2. Participants were coded as responding correctly if they indicated that the puppet who said "Yuck!" was the one who had not tasted the fruits before today. Of the 16 participants who made a choice in the original experiment, 75.00% (n = 12) responded correctly (95% CI: 51.56%-100%). These results suggest that children can use knowledge about the actual subjective rewards that different agents obtain to infer which agent is more likely to have been naïve in her estimate of the expected rewards. However, four children answered incorrectly and two were excluded from analysis and replaced for failing to answer the test question. Thus, to ensure the validity of our interpretation we replicated the experiment. Out of the 16 participants who made a choice in the replication, 75% (n=12) responded correctly (95% CI: 51.56%-100%). Together, these experiments suggests that children can in fact use knowledge about subjective rewards to infer knowledge, and it provides some suggestive evidence that children may find it easier to use information about agent's knowledge to predict their subjective rewards than to use information about agent's rewards to recover information about their unobservable mental states.

Experiment 4

In Experiment 4 we invert the question asked in Experiment 2. Here we see if children can infer which of two agents is more likely to be naive when one shows stable preferences and one does not.

Methods

Participants 16 participants (mean age (SD): 5.64 years (244 days), range 4.04-5.93 years) were recruited at an urban children's museum. Four additional children were tested but excluded from the study because they failed to respond to the inclusion question correctly. See Results.

Stimuli The stimuli were identical to those used in Experiment 1.

Procedure The procedure was identical to Experiment 3 except as follows: First, children were never given any information about whether the puppets said "Yum!" or "Yuck!" after tasting the fruits. Instead, after taking a bite from their fruit, the experimenter said, "Anne kept eating the rambutan. Sally changed her mind and said she wanted an African cucumber instead." As in Experiment 3, at test children were asked, "Can you tell me, which one of our friends didn't know what rambutans tasted like until today?" (Actual fruits counterbalanced throughout.)

Results and Discussion

All results were coded in the same way as Experiments 1-3. Four children failed to respond to the inclusion question correctly and were therefore excluded from analysis and replaced. Children's responses were coded as responding correctly if they indicated that the puppet who changed her mind was the one who had never tasted the fruits before. Of the 16 participants who made a choice, 81.25% (n = 13) responded correctly (95% CI: 58.34-100%).

Together with Experiment 2, these results suggest children understand that relative to naïve agents, knowledgeable agents are more likely to stick to their choices. Moreover, children can infer which agents are more likely to be knowledgeable based on the stability of these choices.

General Discussion

Across four experiments, we studied children's understanding of the relationship between agents' knowledge of their desires and the outcome of these agents' actions. Our results suggest that four and five-year-olds understand that, relative to naïve agents, knowledgeable agents are more likely to obtain high rewards (Experiment 1) and more likely to make stable choices (Experiment 2). Similarly, children believe that agents who obtain high rewards and agents who make stable choices are more likely to be knowledgeable (Experiments 3 and 4 respectively). Collectively, these results suggest that children understand that an agent's choices are not always aligned with the highest utility, but rather with the highest expected utility. As such, agents who have more uncertainty about the value of a target are more likely to obtain a low reward, and more likely to explore different alternatives.

Children's responses in our task could have been driven by their expectations about knowledgeable agents, naïve agents, or both. Future research might see which of these expectations underlies children's reasoning. Additionally, we have emphasized the possibility that children should infer that naïve agents are more likely than knowledgeable agents to make unrewarding choices, and this interpretation is consistent with the results of Experiments 1 and 3. However, children may also believe that, relative to knowledgeable agents, naïve agents are more likely to find exploration rewarding; this kind of reasoning could contribute to the results of Experiments 2 and 4

Computationally, children's intuitions may stem from a categorical distinction between knowledgeable and naïve agents, or from a continuous representation of how the amount of knowledge an agent has influences the quality of their choices. Although these two accounts make similar predictions in our task, the latter theory is more powerful, as it enables observers to reason about intermediate stages of knowledge. Further work is needed to establish exactly how children represent an agent's uncertainty. Furthermore, it is an open question how children represent a goal's reward. For instance, children might assume that each goal has a fixed reward value, which the agent may not know. Alternatively, children may understand that the same outcome can have variable rewards over time (an agent may find apples very rewarding when she's hungry and less so when she's full). Future research might investigate the precise representations underlying children's calculations of agents' utilities.

In these studies, we focused on agents' estimates of the expected reward of their actions. However, the same logic applies to agents' understanding of the cost of actions. For example, Sally might be eager (or reluctant) to run a marathon. However, if you know she does not understand the costs involved, you might not be confident that her current actions will be informative about her future ones. The converse inferences also hold. If you see Sally sign up for a committee and then fail to attend, you might infer that she had not accurately estimated the commitment involved. Recent work suggests that young children are sensitive to the cost of action when reasoning about an agent's preferences or motivation (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Jara-Ettinger, Tenenbaum, & Schulz, 2015). Future work might investigate whether four and five-year-olds also have strong intuitions about how agents' knowledge about the costs of action influences their preferences and choices.

The current results suggest the importance of a more nuanced approach to both empirical and computational work on theory of mind. Past research has focused on learners' ability to draw inferences connecting agents' beliefs, desires, and actions (e.g., Wimmer & Perner, 1983; Baker, Saxe, & Tenenbaum, 2011). Almost universally however, beliefs and desires have been treated as independent, non-interacting variables, and the content of beliefs has been restricted to information about the world. The current study suggests that children understand that agents have beliefs not only about the world, but also about their own preferences. Children understand that as agents gain knowledge about the world, their preferences can change as well. As scientists, we can use this understanding to develop more sophisticated approaches to understanding theory of mind.

Acknowledgments

We thank the Boston Children's Museum and the families who volunteered to participate. We thank Eileen Rivera and Allison Kaslow for help with recruitment, data collection, and video coding. We are grateful to three anonymous reviewers and to Rachel Magid for useful comments. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216, and by the Simons Center for the Social Brain.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-second annual conference of the cognitive science society* (pp. 2469-2474).

Cohen, J. (1994). The earth is round (p<.05). *American psychologist*, *49*(12), 997.

Cumming, G. (2013). The new statistics why and how. *Psychological science*, 0956797613504966.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Gopnik, A., & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child development*, *62*(1), 98-110.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. The MIT Press.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science*, *317*(5843), 1360-1366.

Jara-Ettinger, J., Baker, C. L., & Tenenbaum, J. B. (2012). Learning what is where from social observations. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 515-520).

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015) Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science.*

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015) Children's understanding of the costs and rewards underlying rational action. *Cognition*.

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, *121*(3), 289-298.

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

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, *33*(1), 12.

Wellman, H. M. (1990). *The child's theory of mind* (Vol. 37). Cambridge, MA: MIT press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta‐analysis of theory‐of‐mind development: the truth about false belief. *Child development*, *72*(3), 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.

Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). How infants make sense of intentional action. *Intentions and intentionality: Foundations of social cognition*, 149-169.