The Signature of All Things: Children Infer Knowledge States from Static Images

Madeline Pelz[1] **(mpelz@mit.edu), Laura E. Schulz**[1] **(lschulz@mit.edu),**
Julian Jara-Ettinger[2] **(julian.jara-ettinger@yale.edu)**
[1] Dept. of Brain and Cognitive Sciences, MIT. [2] Dept. of Psychology, Yale University.

Abstract

From minimal observable action, humans make fast, intuitive judgments about what other people think, want, and feel (Heider & Simmel, 1944). Even when no agent is visible, children can infer the presence of intentional agents based on the environmental traces that only agents could leave behind (Saxe et al., 2005; Newman et al., 2010). Here we show that, beyond inferring the presence of agents, four- to six-year-olds can also determine the mental states that best explain an environmental trace. Participants ($N = 35$, M: 5.6 years, range:$4.0-6.8$ years) saw pairs of dresser drawers with different numbers and orientations of open drawers, and they were asked to determine which static scenes was generated by an agent with a given knowledge state (whether the agent wasn't searching at all but was just playing, knew exactly where an object was hidden, knew the approximate location, had no idea where it was hidden, or at first didn't know and then remembered). We compare children's performance to a computational model that extends models of mental-state attribution to consider cases where the behavior is not observed but must be inferred from the structure of the environment. We find that children's graded pattern of responded shows quantitative similarity to the predictions made by our model.

Keywords: cognitive development; theory of mind; computational cognition

Introduction

From simple shapes moving around in a two-dimensional space, adults can draw complex inferences about the mental states and social interactions of agents (Heider & Simmel, 1944). Even without any motion information, adults can make the same kinds of inferences from static scenes. When animal trackers see footprints in the snow, they can infer the type of animal that left them, its likely goal or destination, and sometimes even its age and physical condition; from intersecting tracks, observers may even be able to reconstruct social interactions among many different animals. Inferences like these depend on many cognitive capacities. In order to work backwards from traces in the environment to judgments about mental states, people need to understand the generative process that gave rise to the evidence. This requires understanding that 1. agents have psychological states like goals, desires, and beliefs; 2. that these mental states guide how agents act; and 3. that these actions leave observable changes in the structure of the environment.

As we review below, considerable work suggests that many of these capacities are present early in development. However, the vast majority of work in the development of theory of mind has looked at children's ability to recover mental states from observations of agents' actions; fewer studies have looked at children's ability to infer the presence of an agent when no agent is present. How general and flexible is our ability to recover information about agents from information in the environment? Can children draw inferences about agents from static scenes? Are they able to go beyond inferring just the presence of agents to inferring their behavior, goals, and mental states? We focus on four- to six-year-olds, an age at which children can recover a wide range of mental states—desires, ignorance, knowledge, and false beliefs—from observed behavior (Wellman, 2014). We ask whether preschoolers can make similarly rich inferences when neither the agent nor the behavior is observed but must be recovered from static traces. To interpret these results, we also present a computational model that extends classical models of mental-state attribution (C. L. Baker et al., 2017; Jara-Ettinger et al., 2019; Lucas et al., 2014; Jern et al., 2017) to infer mental states from indirect environmental traces.

Even infants interpret other people's behavior as goal-directed (Sommerville et al., 2005; Woodward, 1998), and they assume that these goals are the result of agents trying to maximize the subjective rewards they obtain, while minimizing the costs that they incur (Csibra et al., 2003; Jara-Ettinger et al., 2016; Liu et al., 2017; Lucas et al., 2014). Through these assumptions, children can draw rich inferences about agents' mental states, such as their preferences (Pesowski et al., 2016; Jara-Ettinger et al., 2015) and knowledge (Jara-Ettinger et al., 2017); whether false belief reasoning inferences emerge before four and five remains very controversial (see e.g., Baillargeon et al., 2010; Powell et al., 2018). Indeed, even in four- and five-year-olds, some kinds of false belief reasoning are relatively fragile (e.g., Bradmetz & Schneider, 1999; Hogrefe et al., 1986).

Beyond a capacity to infer mental states from observable behavior, young children can also infer the presence of agents from indirect evidence. Infants and toddlers infer that a hand, but not an inanimate, object is hidden behind a screen when objects appear to move spontaneously (e.g., when an otherwise stationary objects flies through the air; Saxe et al. 2005) or behave probabilistically (e.g., a mechanical lever switches from producing one effect to producing another; Wu et al. 2016). Similarly, infants expect a hand rather than a mechanical claw when they see an improbable versus random sample drawn from a population (Ma & Xu, 2013), and infer that an intentional agent rather than a physical force constructed an orderly versus scattered array of blocks (Newman et al., 2010). Children also expect that, while non-agentive forces always increase entropy in the environment, agents can increase or decrease it at will (Friedman, 2001), and they selectively infer that an agent is present only when they hear ordered (but not random) sequences of sounds that cannot be explained away by the environment (Schachner & Kim, 2018).

The studies above have looked at children's inferences given dynamic sequences of events (e.g., moving objects; sequential tones, etc.) and looked only at whether children infer the presence of agents versus non-agents. To our knowledge, no work has looked at the inferences children draw from static scenes, or whether children can recover agents' behavior, goals, and mental states from the structure of the environment. Inspired by formal work on theory of mind, here we look at whether children can use relatively nuanced differences in the physical structure of a scene to recover agents' actions, goals, knowledge, and potentially even false beliefs.

Computational Framework

Method

Our computational model is motivated by a growing body of work showing that we intuitively expects agents to maximize their subjective utilities—the difference between the costs they incur and the rewards they obtain (Jara-Ettinger et al., 2016; Lucas et al., 2014; Jern et al., 2017). Using this assumption, inferences about other people's mental states can be modeled as a form of *inverse planning*, where observers 'invert' agents' decision-making process to recover the mental states that best explain their observed behavior (C. L. Baker et al., 2017, 2009; Jara-Ettinger et al., 2019). Our model builds on this framework, but diverges in that we consider cases where observers cannot directly see the actions an agent took, and must instead infer them from an indirect environmental trace.

For simplicity, we explain the logic of our model in the context of our experiment. Imagine encountering a dresser like the ones shown in Figure 1, with the knowledge that someone had been searching for their sweater. You can see the drawers that they opened, but not the order in which they carried out the search. Intuitively, seeings these static scenes reveals what agents' might have known when they began searching for the sweater.

Formally, we define a knowledge state K as a probability distribution over possible drawers. This allows us to capture a wide range of possible knowledge states from full ignorance (corresponding to a uniform distribution) to perfect knowledge (corresponding a distribution where the drawer with the sweater has probability 1, and all other drawers have probability 0). Given an observed dresser in state D, the probability that an agent had knowledge state K is given by

$$p(K|D) \propto p(D|K)p(K) \qquad (1)$$

where $p(K)$ is the prior distribution over what the agent knew, and $P(D|K)$ is the likelihood that an agent with knowledge state K would produce the environmental trace seen in the dresser's state D. To compute this likelihood, we consider the potential unobservable actions that the agent may have taken, such that

$$p(D|K) = \sum_{t \in T} p(D|t)p(t|K) \qquad (2)$$

where t is a trajectory opening different drawers (from the set T of all possible trajectories). $p(D|t)$ is the probability that the dresser would end in state D if the agent took actions t, and $p(t|K)$ is the probability that the agent would choose to take actions t given knowledge K. Thus, our model can be thought of as decomposing the inference problem into two components: an understanding of how mental states generate actions (captured in $p(t|K)$), and an understanding of how actions produce environmental traces (captured in $p(D|t)$).

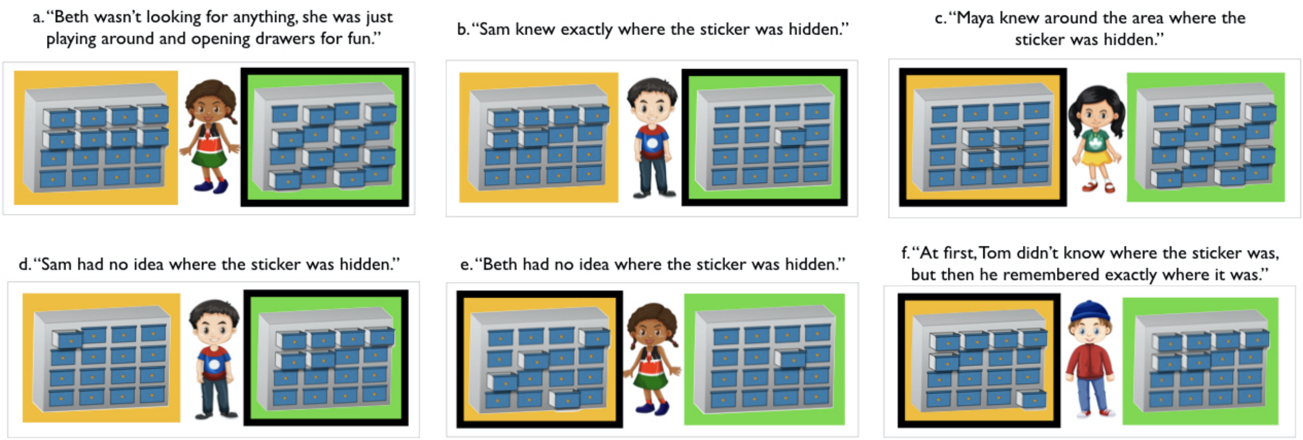
Table 1: Hypotheses and Abbreviations

Play	He wasn't searching for anything, he was just opening drawers to make a design
Exact	He knew exactly where the item was hidden
Approx	He knew the approximate location of the item
None	He had no idea where the item was hidden
Recall	At first, he didn't know where it was hidden, but then he remembered

Here we define the probability of generating different environmental traces ($p(D|t)$) as 1 when the trajectory opens all and only the drawers that are open in D and 0 otherwise. To compute the probability that an agent would take trajectory t given their knowledge state ($p(t|k)$), we use a Partially Observable Markov Decision Process (POMDP; C. L. Baker et al., 2017).

POMDPs are a general planning framework to determine how agents with incomplete knowledge about the world act to fulfill their goals (encoded as a reward function over possible states of the world). In POMDPs, an agent can take actions (corresponding to opening different drawers in our paradigm) that both reveal information (in our case, whether the drawer was empty or not) and can also change the world state (such as leaving an environmental trace). Within the POMDP framework, it is possible to compute the sequences of actions that maximize an agent's reward function, thus capturing the expectation of how agents act to maximize the rewards while minimizing the costs they incur (Csibra 2003; Jara-Ettinger et al. 2016; C. Baker 2004).

POMDPs are designed to produce an optimal solution. In our paradigm, that would mean that ($p(t|k)$) would be positive only when the trajectory t is an optimal search for the sweater, given knowledge k. In our model we relax this assumption, using a probabilistic POMDP that assumes that trajectories with higher utilities have a higher probability of being generated (without expecting that agents are strictly optimal). We achieve this using the standard approach of soft-maxing the value function (see C. L. Baker et al. 2017 for extended discussion of probabilistic POMDPs in the context of Theory of Mind models).


a. "Beth wasn't looking for anything, she was just playing around and opening drawers for fun."

b. "Sam knew exactly where the sticker was hidden."

c. "Maya knew around the area where the sticker was hidden."

d. "Sam had no idea where the sticker was hidden."

e. "Beth had no idea where the sticker was hidden."

f. "At first, Tom didn't know where the sticker was, but then he remembered exactly where it was."

Figure 1: Stimuli used in the child behavior experiment. Each forced-choice question showed an image of a character with a certain knowledge state, a target image of the predicted pattern of drawers (highlighted by black outline), and a distractor.

While POMDPs are a useful way of capturing the expectation that agents are efficient, they can also be overly restrictive. Intuitively, people have a naive expectation that agents are more likely to begin searching on the top-left drawer, and that they are biased in searching from left to right. Because these expectations cannot be explained by appealing to efficiency in space, we modified our POMDP to assume that agents are more likely to search from left to right, usually starting on the top-left drawer.

Our model considered five different hypothesis, shown in Table 1. We began by considering a first hypotheses where knowledge is irrelevant. In the *Play* hypothesis, we assumed that the agent was not searching for their sweater. Because in this case, the agent is not expected to be fulfilling a goal efficiently, we estimated the posterior probability of play by placing a uniform likelihood over all trajectories in Eq. 2. The next four rows of Table 1 show the basic types of knowledge hypotheses that we considered. *Exact* hypotheses are those were a single drawer has probability 1 (16 hypotheses total), *Approximate* hypotheses are those where the agent recalls the approximate location of the sweater (using a decaying probability distribution centered at the intersection of any four drawers; 9 hypotheses), *None* hypotheses are those where all drawers have the same probability (1 hypothesis), and *Recall* hypotheses are those where the agent was initially ignorant and then suddenly remembers its exact location.

Child Behavior Experiment

Participants

In accordance with an Open Science Framework preregistration, 35 children (mean: 5.6 years, range: 4.0-6.8 years) were recruited from an urban children's museum. Four additional children were tested but not included in the sample; two for parental or sibling interference, one for responding before the prompt was completed, and one that chose not to complete the task.

Methods

Children first completed a warm-up task where they counted the drawers in a small (27cm x 22cm) set of 16 drawers arranged in a 4x4 grid. They were then asked to close their eyes while the experimenter hid a sticker in one of the 16 drawers, and then were given as much time as needed to find the sticker. Once children were familiar with the concept of searching within the drawers, they then moved to the test phase.

Children were told they would see a picture of a character who had searched for a sticker in the same game that they had just played. The experimenter then explained the task to the child, saying "Remember when you were looking for your sticker? You had no idea where I hid the sticker in your game, because your eyes were closed. But now I'm going to show you some friends who were playing this game before, and the characters you're going to see might have known something different about where their sticker was hidden. For each character, I'm going to tell you what they knew about where their sticker was hidden, and show you two different sets of drawers that they could have used for their game. It's your job to listen to what they knew about where their sticker was, and to tell me which set of drawers you think they looked inside for their sticker."

Participants were then shown six forced-choice test questions presented in a randomized order, each with an image of a character, a target image of the predicted answer, and a distractor (Figure 1. As a proof of concept, for this experiment we hand-matched targets and distractors to capture our intuition about what might be the clearest discriminations. Half the items used as targets were also used as distractors. For each trial, the experimenter directed the child's attention to one set of drawers and then the other. They were then told what this character knew about where their sticker was hidden (see Figure 1. for stimuli and prompts), and then asked "Which set of drawers do you think [insert name] was looking inside of?" Children indicated their choice by naming the color of the box around the set of drawers, or by pointing. This procedure was repeated for each of the six trials, and children did not receive feedback between trials.

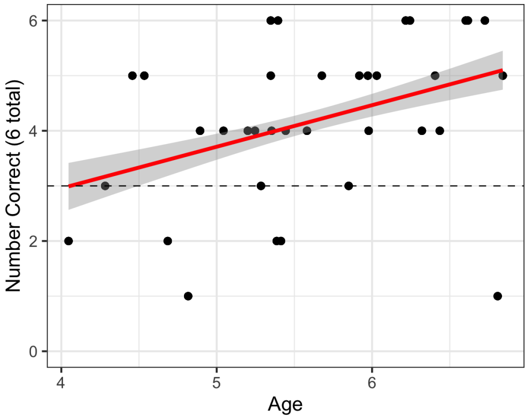


Figure 2: Overall, children performed above chance, answering correctly in 70% of trials. Each point represents a participant with their age on the x-axis and the number of correct trials on the y axis. The horizontal dotted line shows expected chance performance (three correct answers) and the red line shows the best linear fit with 95% confidence band in light gray.

Results

When collapsed across all 6 trials (Figure 2.), children performed above chance, succeeding on an average of 4.17 trials (95% confidence intervals: $[3.67, 4.67]$, $p < 0.0001$, by one sample t-test). In addition, a pre-registered analysis uncovered a significant age effect, with children improving at the task across age ($\beta = .75$, $R^2 = .15$, $p < 0.0001$; Figure 2).

When split by question (Figure 3A.), children performed significantly above chance in the play, approximate, and efficient trials, (95% confidence intervals = $[0.60-0.90]$, $p = 0.002$, $[0.57-0.89]$, $p = 0.006$, $[0.57-0.89]$, $p = 0.006$, respectively), but not in the exact, random, or remembering trials (95% confidence intervals $[0.50-0.81]$, $p = 0.089$, $[0.45-0.79]$, $p = 0.175$), $[0.45-0.79]$, $p = 0.175$, respectively.)

Model Results

To generate model predictions, we used a softmax parameter of $\tau = 1.5$. Our model outputs a posterior distribution over every possible knowledge state, given a dresser. To translate these inferences into task judgments, we calculated the normalized probability that an agent would generate the environmental trace of each pair of dresser, given the target knowledge state. For instance, on the trial in Figure 1B, we extracted the posterior probability that only the first dresser was generated by an agent with no knowledge, and the posterior probability that only the second dresser was generated by an agent with no knowledge, and we then normalized these judgments so that they summed to 1.

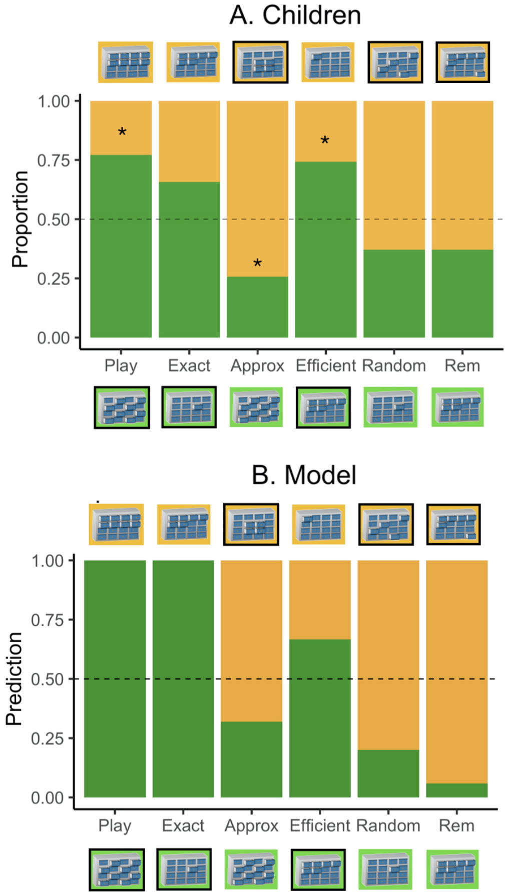


Figure 3: A. Behavioral results as a function of question type. When split by question, children performed significantly above chance in the play, approximate, and efficient trials, but not in the exact, remembering, or random search trials. Each bar column shows the distribution of responses for each trial type. B. Model predictions for children's behavioral test trials. Each bar shows the probabilities assigned to the target and distractor image in each pair, with colors indicating its preferred image for each tested trial.

Overall, the model aligned well with preschoolers' responses, with an overall correlation of 0.856 (95% confidence intervals = $[0.147-0.984]$, $p = 0.029$). The model had the strongest preferences in the play, exact, and remember conditions, preferring the target image with a probability of 1, 1, and 0.941, respectively. Although preschoolers also performed well above chance in these conditions, their responses were, unsurprisingly, noisier. In the approximate and random cases, the model made predictions that were qualitatively similar to children's estimates, although children did not significantly prefer the target dresser in the random case. However, children drew stronger inferences than the model about the pattern consistent with an ignorant agent who engaged in efficient search (i.e., searching adjacent drawers and finding the target by chance); the model only preferred the target pattern with a probability of 0.667. This is because our model had a strong expectation for rational action, and an expectation that agents begin searching on the top-left drawer. Thus, opening the top-left drawer already suggested that the agent had no additional initial knowledge to go from, and just began searching in the most convenient drawer. From this standpoint, the model intuitively reasoned that both cases were consistent with the agent having no knowledge, and that the agent just got lucky in the left display by finding it on the first try.

General Discussion

In the current study, we found that four- to six-year-olds were able to match a range of mental states with images of dresser drawers, given only the information that an agent was looking for a target and the number and location of the open drawers. Children's judgments were consistent with a computational framework that performs inverse planning to infer the order in which the drawers were searched, as well as the knowledge state that best explains the observed evidence.

Children performed above chance on the play, approximate, and efficient trials, but not on the exact, random, or remembering trials. With regard to the random condition, we believe the model and the children might both have accurately represented information that our experimental intuitions (in hand-picking targets and distractors) missed. We tested two cases in which the agent was ignorant about the object's location: one in which he searched efficiently, and one in which he searched randomly. In the ignorant efficient search condition, children chose the dresser with more drawers open, suggesting they were sensitive to the fact that it was unlikely that an ignorant agent would find the object in the first drawer he searched. Interestingly however, the children did not privilege this distinction in the random search case. We suggest this might be because both options were unlikely: It is unlikely that an ignorant agent would find the object on his first try, but it is also unlikely that an agent would engage in unnecessarily costly actions (an agent who respected principles of rational action would follow a continuous path rather than open non-adjacent drawers). Thus both options may have been similarly improbable.

On the remembering task, the model chose the target consistent with our adult, experimenter intuitions but four- to six-year-olds did not. This is perhaps unsurprising since although children start to pass classic unexpected transfer false belief tasks (Wimmer & Perner, 1983) around age four and five, many tasks involving false beliefs remain challenging well until middle childhood (e.g. Bradmetz & Schneider 1999). "Remembering" is an especially complex belief state, since it involves correcting an initial false belief about one's own ignorance. Inferring that an initially ignorant agent changed into a knowledgeable agent, midway through search, may have been especially difficult for young children, especially in the absence of any agent or observed action at all. We note also that although we framed this as a remembering condition, children might have been more successful if we had framed the task directly in terms of false beliefs. (The target behavior is consistent with remembering the location midway through search but it is also consistent with the agent initially having a false belief that the object was in the lower right corner of the dresser and then realizing he was wrong and having to search from scratch). Indeed, the mere fact that there are two plausible construals may have posed a challenge for children. Future research might disambiguate these accounts.

In the current study, we presented children with a two-alternative forced-choice paradigm with hand-selected comparisons; as a preliminary test of children's abilities, reducing task demands was a priority. But in future work it would be interesting to know if children might succeed in more complex tasks, and whether they might even spontaneously articulate explanations for the drawers' layout invoking the target unobserved mental states. In future work, we also intend to extend this study to adults, where we can vary the task in more quantitatively precise ways. For example, we can ask adults to rank the likelihood of each mental state based on each image and compare their responses to the model across many different patterns of drawers. We can also ask adults for free-response explanations of the pattern of drawers to see how often adults converge on the target hypotheses even when the options are unconstrained.

However, the current findings already suggest the sophistication of children's theory of mind. Extending what has been previously demonstrated about young children's understanding of principles rational action, we found that children can integrate their understanding of agents' mental states with an understanding of the costs and rewards of agents' actions to match what agents did and didn't know from the alterations they made in the structure of the environment. Our study also points to the richness and flexibility of our intuitive psychology. It is unlikely that anyone has ever been asked to match mental inferences with a set of bureau drawers before, and yet four-, five-, and six-year-olds readily recovered approximate knowledge, ignorance, and play from the evidence they observed. While such inferences may seem trivial in the context of bureau drawers, if we return to our original example of observing animal tracks, it is clear why it might be critical to recover the unobserved behavior and mental states of agents responsible for observed changes in the environment. By investigating the computations that underlie this kind of reasoning, we can come closer to understanding the nature of our own minds—and our ability to imagine the minds of others.

References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.

Baker, C. (2004). Functional selectivity of human extrastriate visual cortex at high resolution. *Journal of Vision*, *4*(8), 88.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

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

Bradmetz, J., & Schneider, R. (1999). Is little red riding hood afraid of her grandmother? cognitive vs. emotional response to a false belief. *British Journal of Developmental Psychology*, *17*(4), 501–514.

Csibra, G. (2003, March). Teleological and referential understanding of action in infancy. *Philos Trans R Soc Lond B Biol Sci*, *358*(1431), 447–458.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Friedman, W. J. (2001). The development of an intuitive understanding of entropy. *Child Development*, *72*(2), 460–473.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, *57*(3), 567–582.

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. (2019, Dec). *The naive utility calculus as a unified, quantitative framework for action understanding.* PsyArXiv. Retrieved from psyarxiv.com/e8xsv doi: 10.31234/osf.io/e8xsv

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Ma, L., & Xu, F. (2013). Preverbal infants infer intentional agents from the perception of regularity. *Developmental psychology*, *49*(7), 1330.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., & Wynn, K. (2010). Early understandings of the link between agents and order. *PNAS*.

Pesowski, M. L., Denison, S., & Friedman, O. (2016). Young children infer preferences from a single action, but not if it is constrained. *Cognition*, *155*, 168–175.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50.

Saxe, R., Tenenbaum, J., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001.

Schachner, A., & Kim, M. (2018). Alternative causal explanations for order break the link between order and agents.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1–B11.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

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. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Wu, Y., Muentener, P., & Schulz, L. E. (2016). The invisible hand: Toddlers connect probabilistic events with agentive causes. *Cognitive science*, *40*(8), 1854–1876.